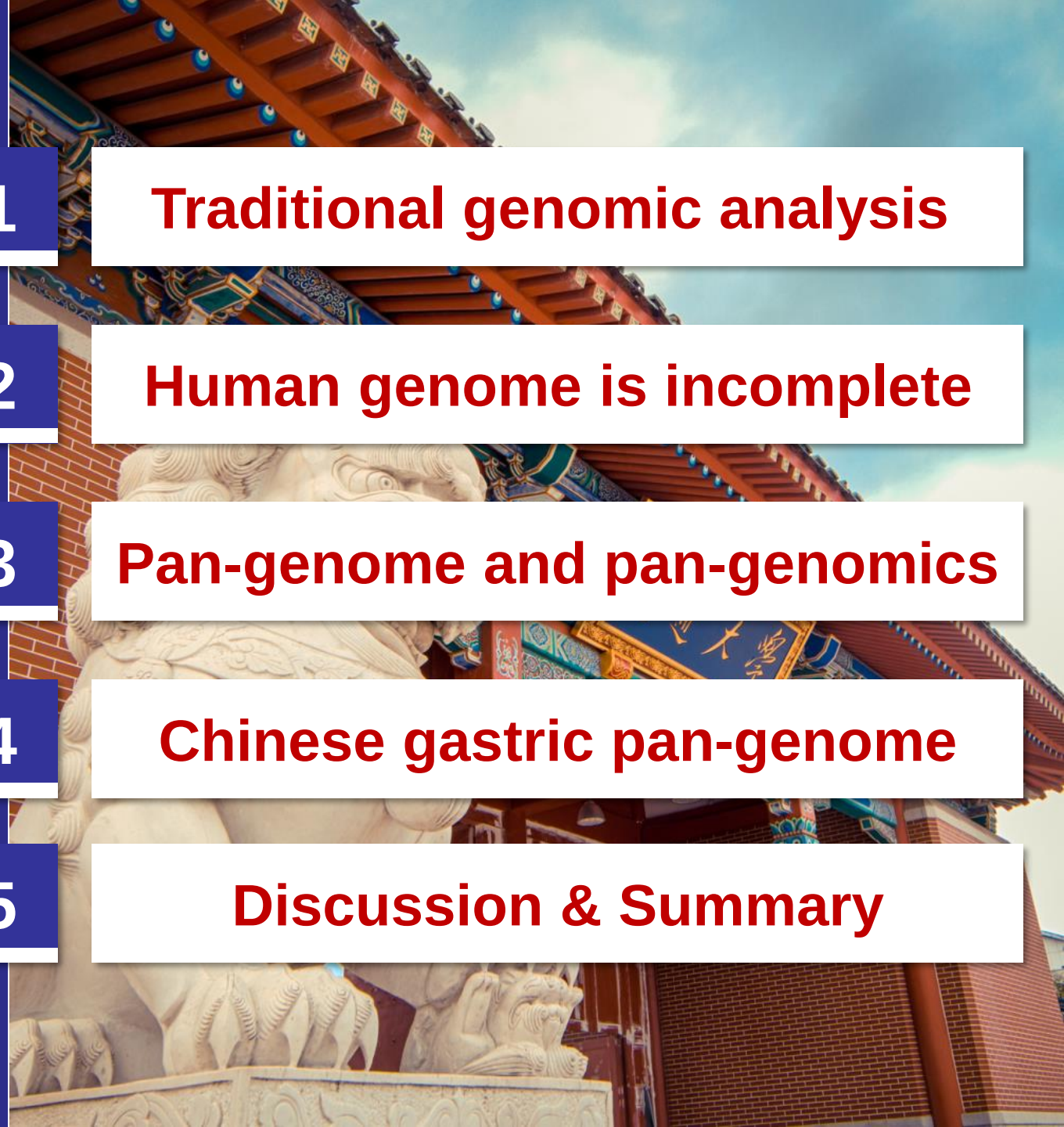
Pan-genomic analysis of Chinese gastric cancer

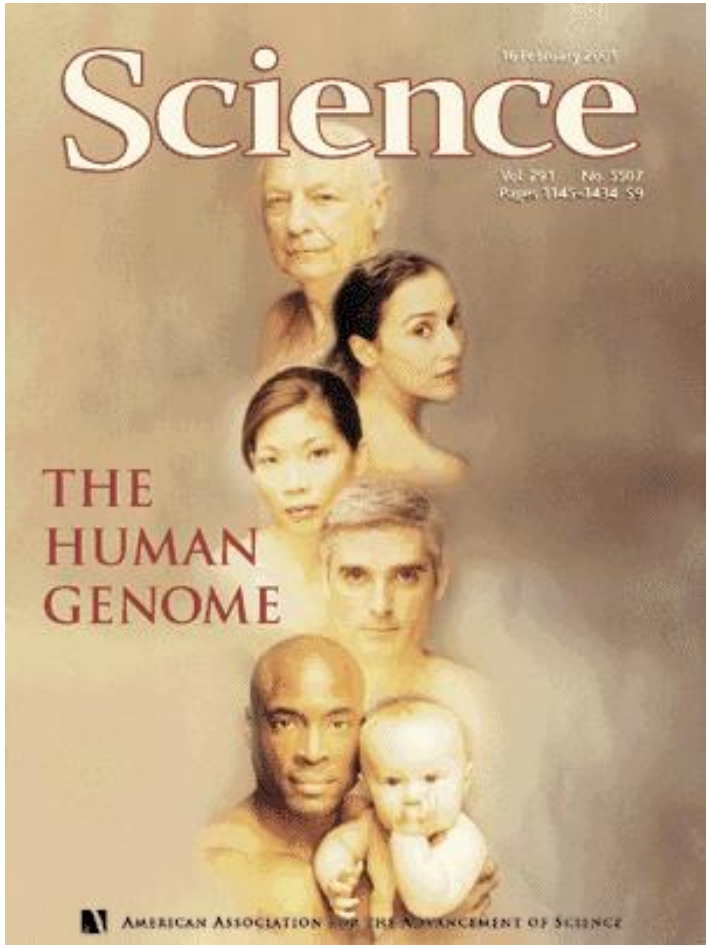
Prof. Chaochun Wei
Department of Bioinformatics & Biostatistics
School of Life Sciences & Biotechnology
Shanghai Jiao Tong University

Contents



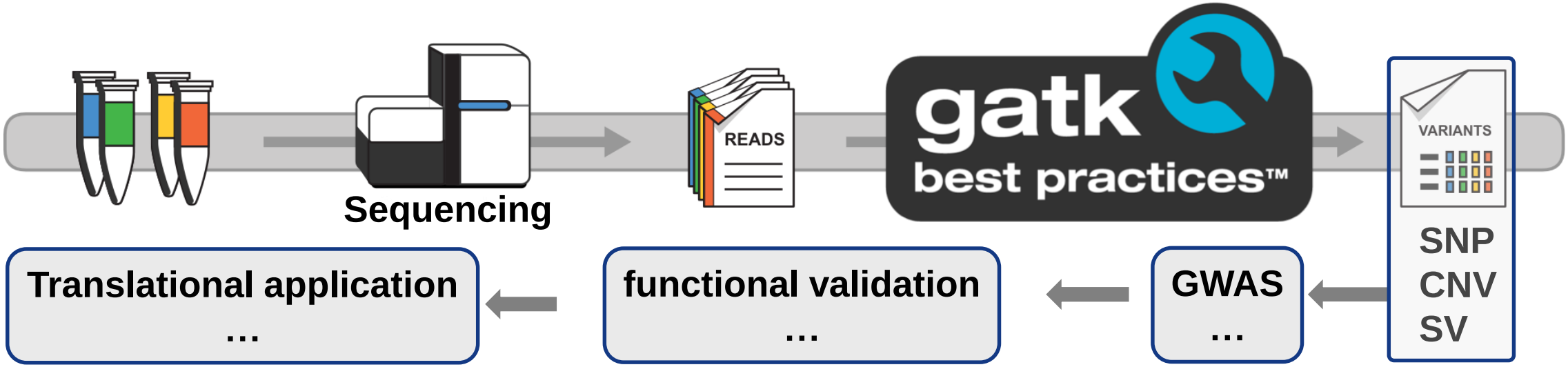
The human genome project (1990-2003)

Feb. 15, 2001 *Nature*



Feb. 16, 2001 *Science*

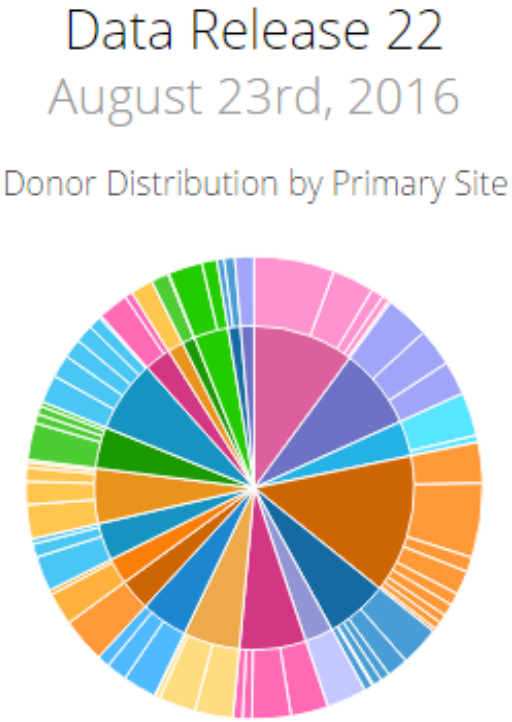
Disease genomics analysis pipeline



https://gatk.broadinstitute.org/hc/en-us

Genome Research, 2010, 20(9): 1297-1303

Caner Genomes
肿瘤基因组



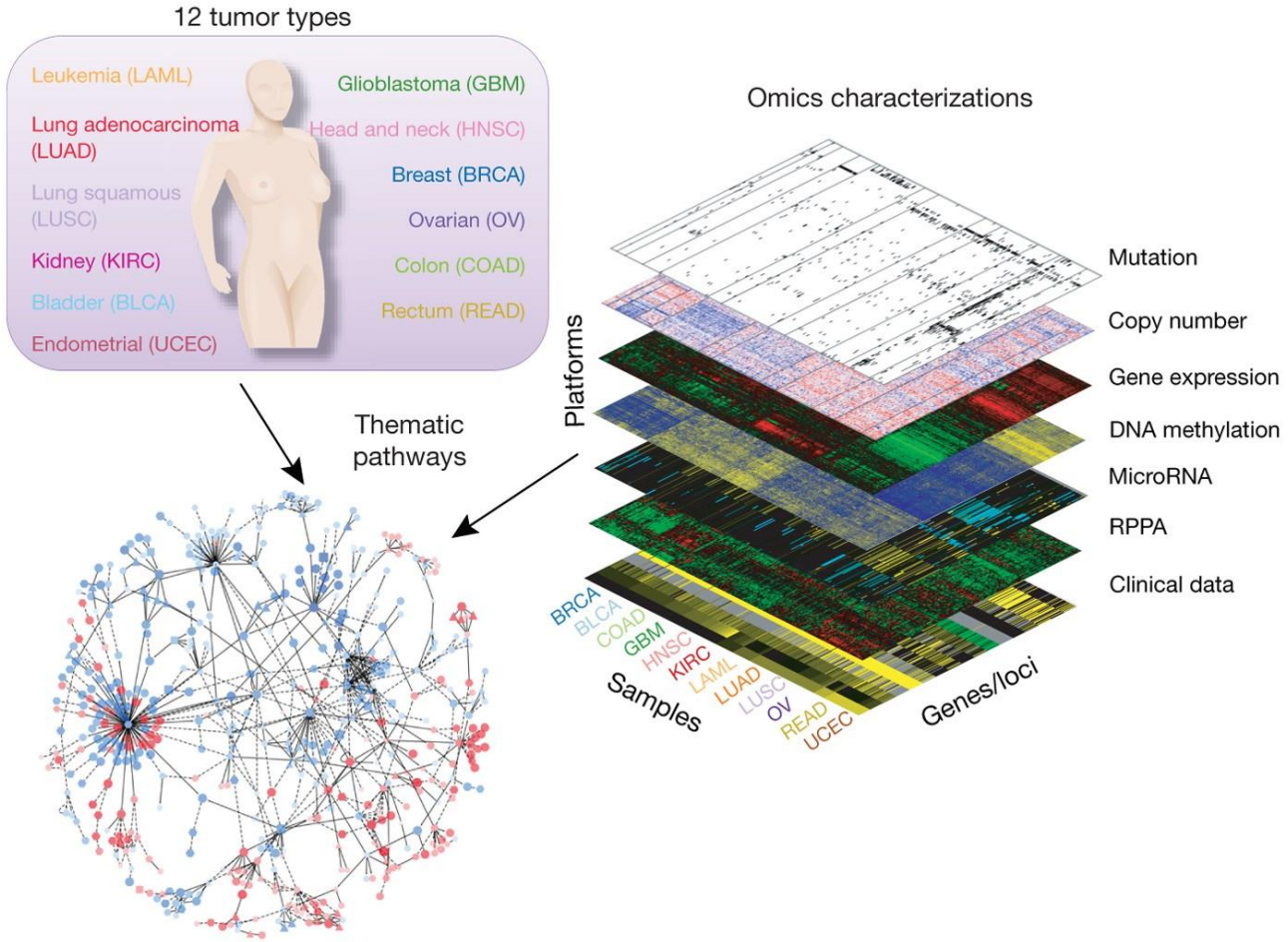
Data release 28
March 27, 2019

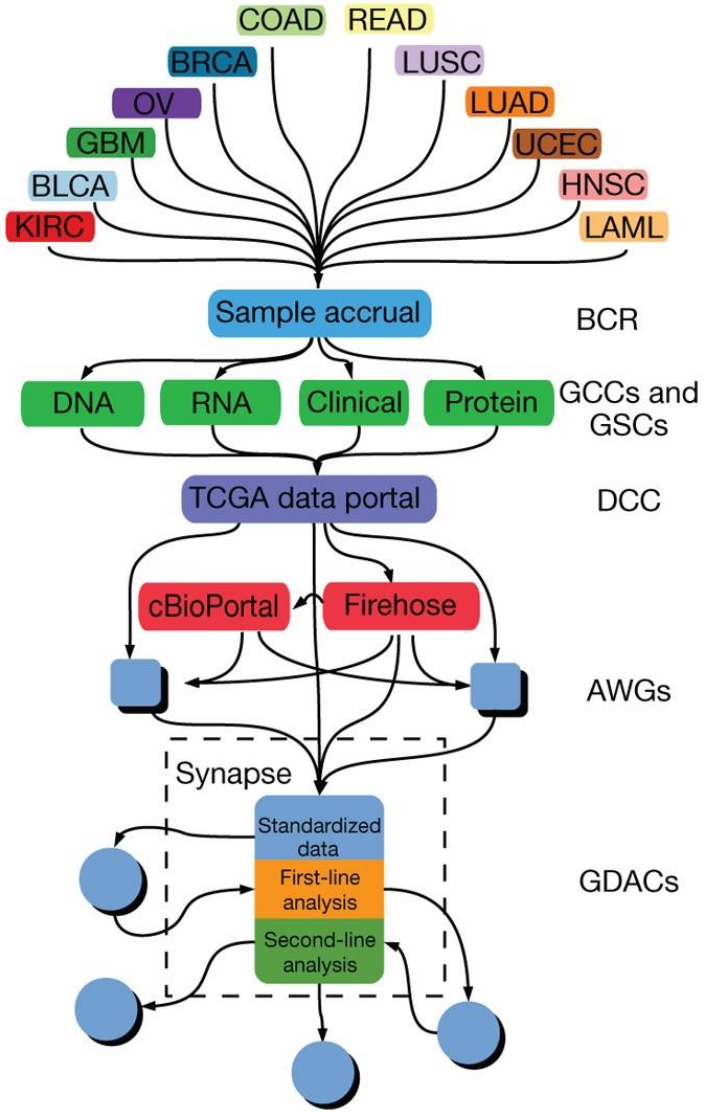
Cancer projects	70	86
Cancer primary sites	21	22
Donors with molecular data in DCC	16,236	22,330
Total Donors	19,290	24,289
Simple somatic mutations	46,429,997	81,782,588
Mutated Genes	57,658	

TCGA：http://cancergenome.nih.gov/

ICGC：http://icgc.org/

TCGA Pan-Cancer project





Nature Genetics, 2013, 45, 1113-1120

Human genome study paradigm: ignoring the non-reference genomic sequences

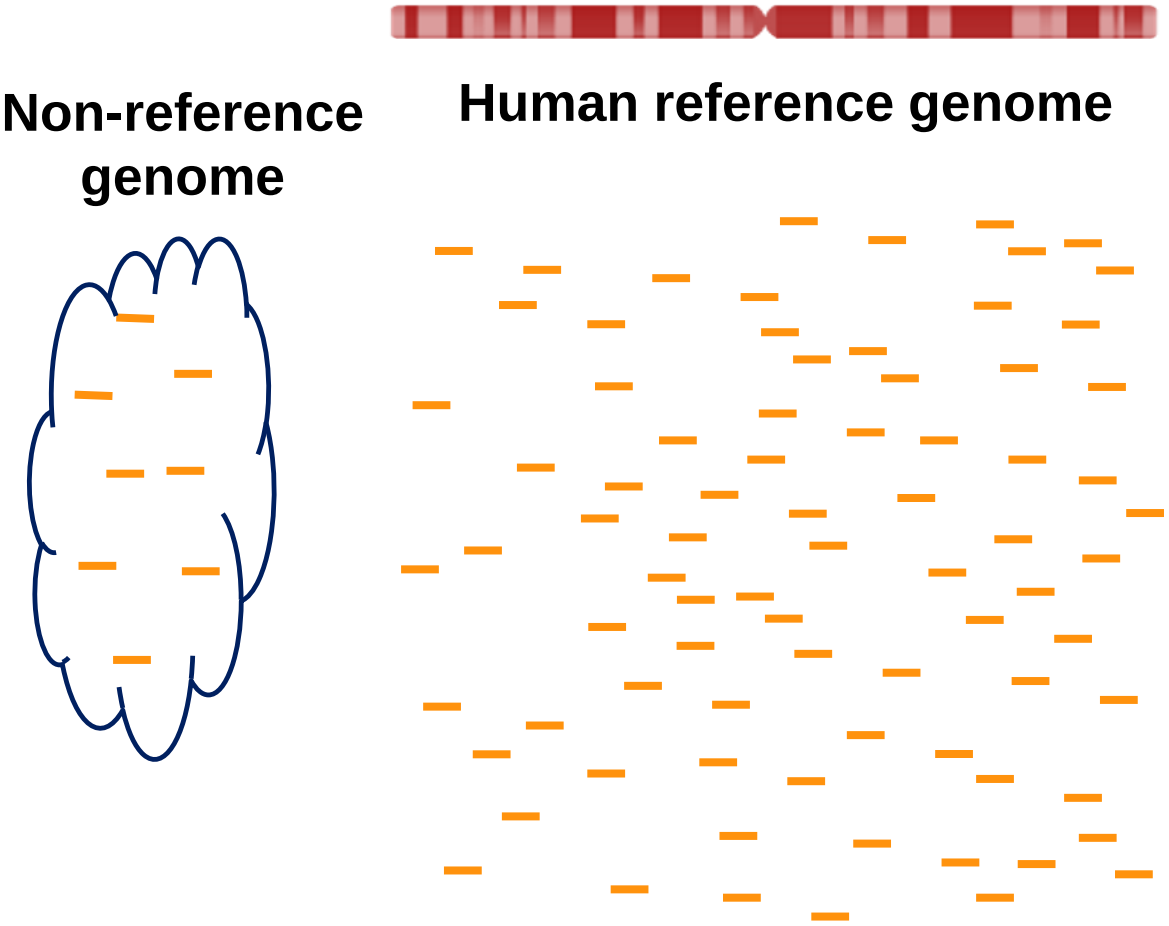
2001 — **Human reference genome**
Nature

2010 2012 — **1000 human Genome**
Nature

2016 — **10,000 human genomes**
Telenti et al, ***PNAS***

2017 — **320,000 Iceland human genomes**
Kehr et al, ***Nature Genetics***

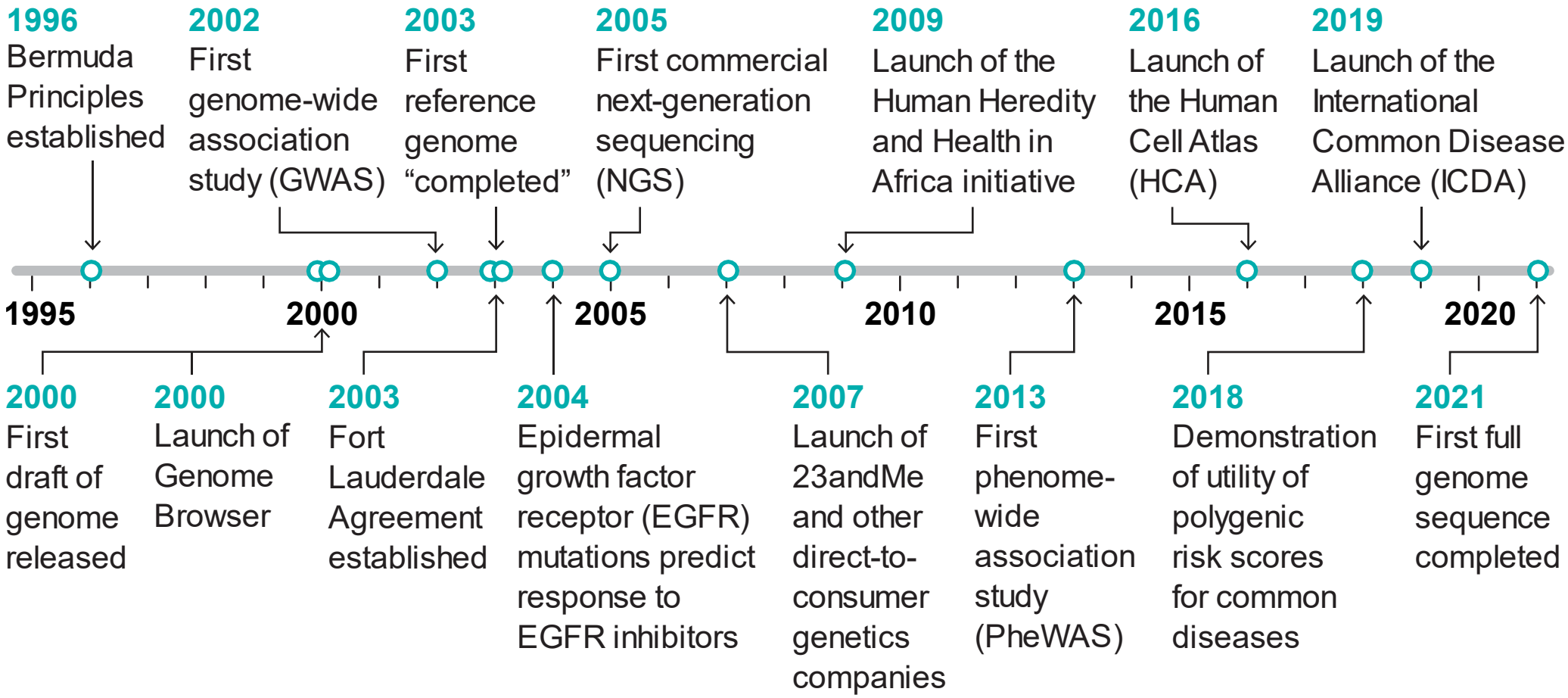
2018 2020 — **Cancer genomes**
Hoadley et al, ***Cell;***
TCGA/ICGC, ***Nature***



2. The human reference genome is incomplete

- sequencing and assembly errors
- underrepresented populations

The legacy of the Human Genome Project



SCIENCE•23 Sep 2021 373(6562) :1442-1443

The first complete human genome, Science, 2021

Nurk, S. et al. Science, April 1, 2022 376(6588):44-53

SPECIAL SECTION COMPLETING THE HUMAN GENOME

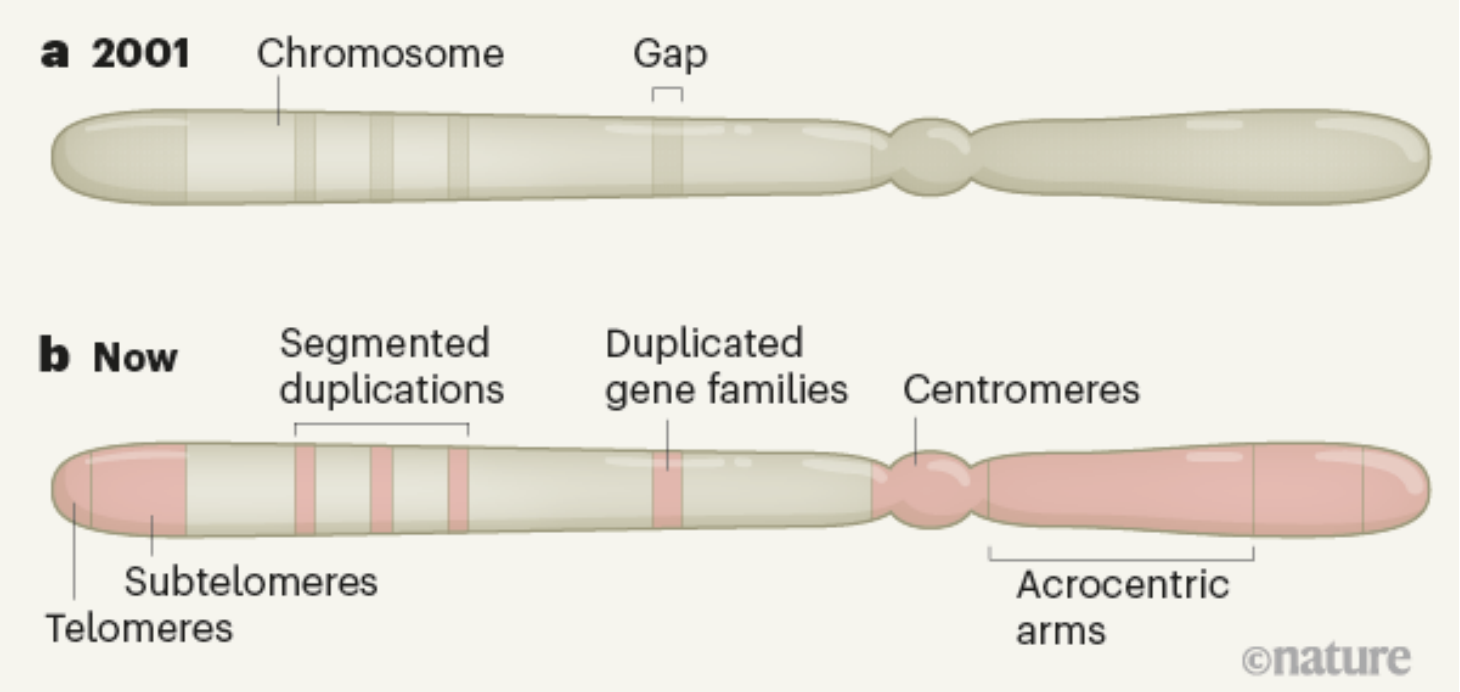
RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29],

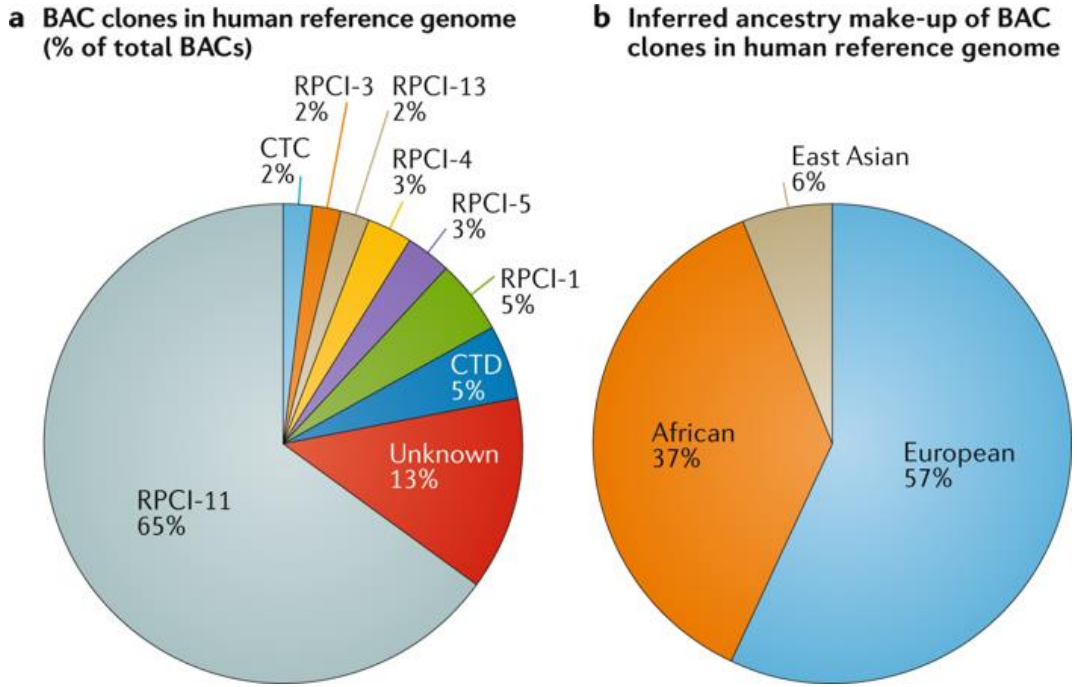
~150Mbp sequences are missing in the human reference genome

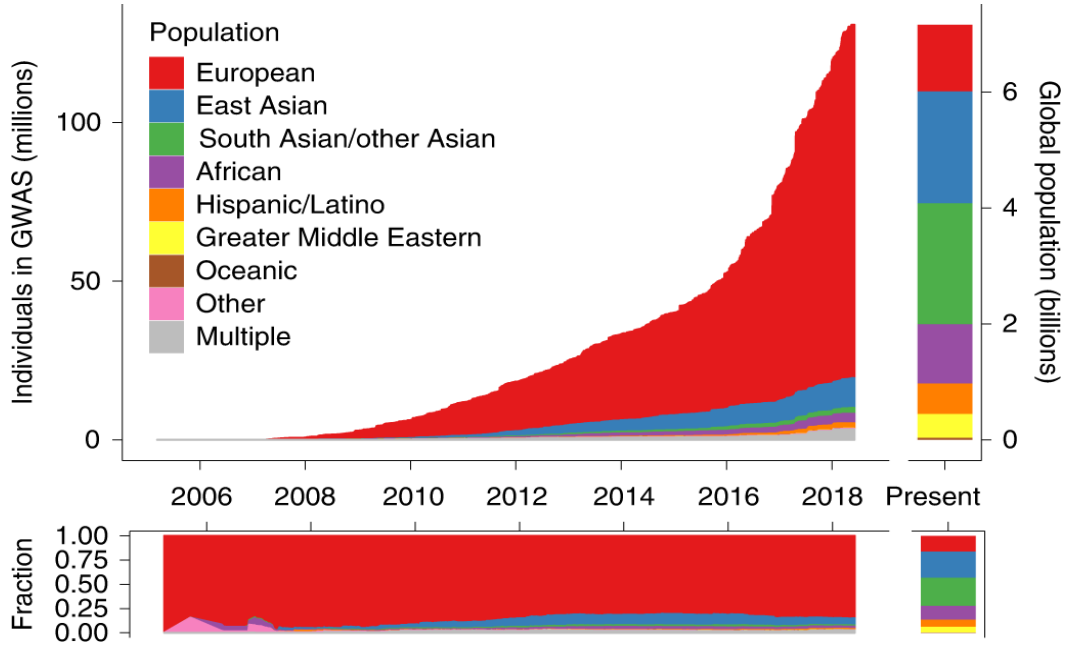


*The sizes of regions are exaggerated in figure

Miga, K. *Nature*, 2021, 590: 217

The human reference genome：Under represented populations





Sample bias in GWAS projects

East Asian：only ~6%

Martin, A.R., et al., *Nature Genetics, 2019*

Sherman and Salzberg，*Nature Reviews Genetics*，2020

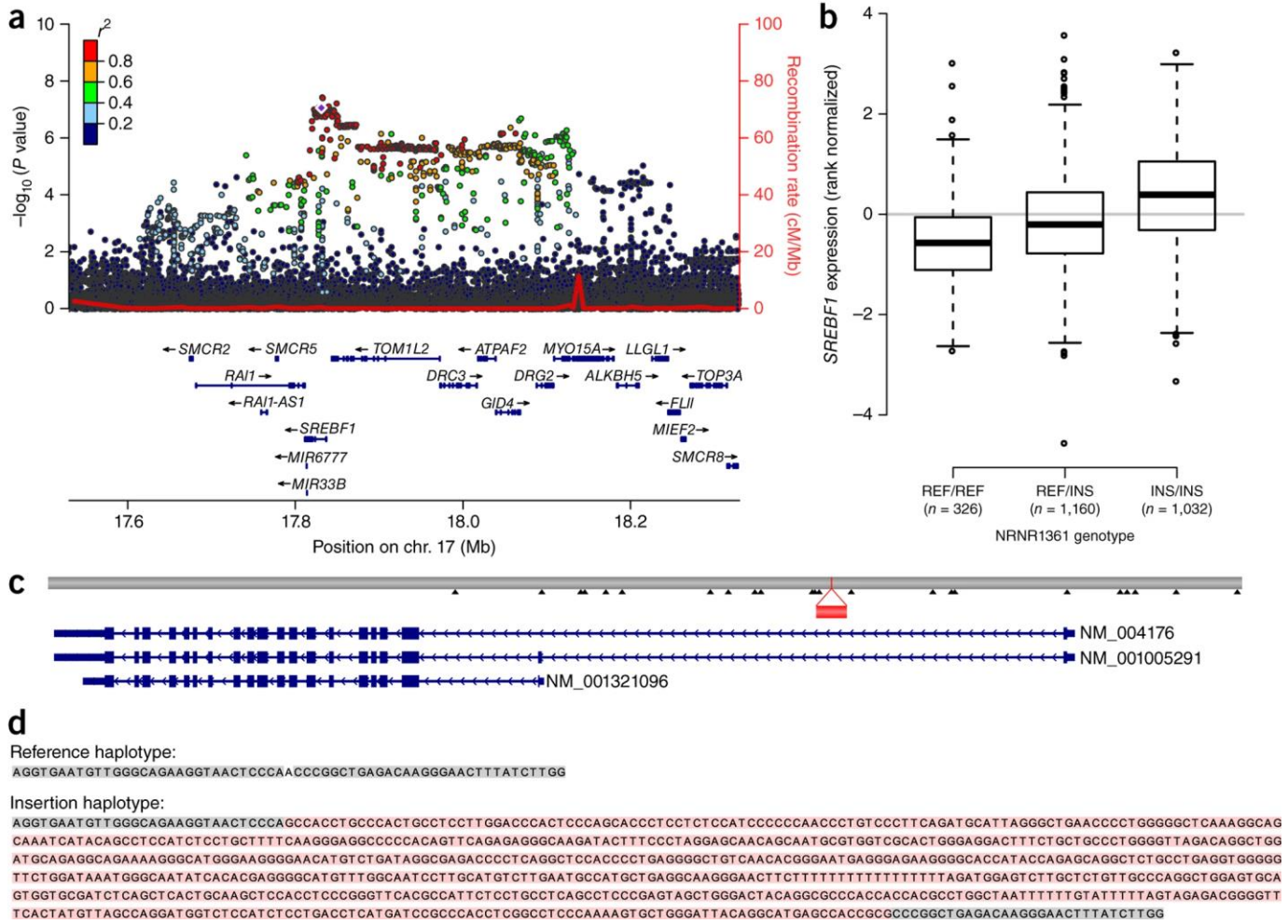
The human reference genome is incomplete!

Population	Samples	Total length of novel seq	Novel sequence per individual	Year	Journal
Swedish	1,000 (subset of 2)	**46 Mb (17.3 Mb)**	0.6 Mb (12.1 Mb)	2019 (2018)	***MBE*, *Genes***
Han Chinese	275	**29.5 Mb**	~5 Mb fully unaligned + ~6 Mb partially unaligned	2019	***Genome Biol.***
Mixed	154	**60 Mb**	14.2 Mb	2019	***Nat. Commun.***
Mixed	15	**21.3 Mb**	6.4 Mb	2019	***Cell***
African	910	**296.5 Mb**	2.5 Mb	2019	***Nat. Genet.***
Mixed, 1KGP	45	**61.6 Mb**	17,700–20,500 insertions	2016	***Human Genetics***

Novel sequences：~1%-10% the size of the human genome

Sherman and Salzberg，*Nature Reviews Genetics*，2020

Non-reference genomic sequences can be important



This 766bp non-reference sequence variant in Icelanders is associated with myocardial infarction!

Kehr et al, *Nature Genetics*, 2017.

TECHNICAL REPORT

https://doi.org/10.1038/s41588-022-01043-w

nature genetics

Check for updates

OPEN

Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes

Annu. Rev. Genom. Hum. Genet. 2021. 22:81–102

First published as a Review in Advance on April 30, 2021

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-120120-081921

Perspective

The Human Pangenome Project: a global resource to map genomic diversity

https://doi.org/10.1038/s41586-022-04601-8

Received: 30 August 2021

Accepted: 1 March 2022

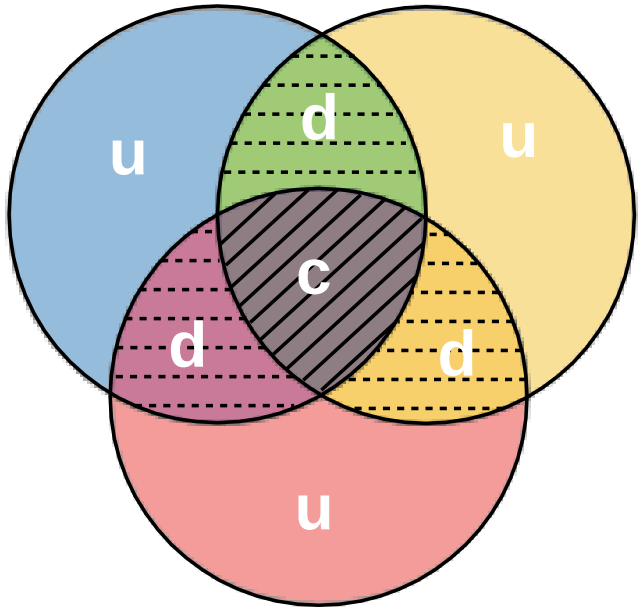
Published online: 20 April 2022

Ting Wang[1,2,3]✉, Lucinda Antonacci-Fulton[3], Kerstin Howe[4], Heather A. Lawson[1], Julian K. Lucas[5], Adam M. Phillippy[6], Alice B. Popejoy[7], Mobin Asri[5], Caryn Carson[1,2,3], Mark J. P. Chaisson[8], Xian Chang[5], Robert Cook-Deegan[9], Adam L. Felsenfeld[10], Robert S. Fulton[3], Erik P. Garrison[11], Nanibaa' A. Garrison[12,13,14], Tina A. Graves-Lindsay[3], Hanlee Ji[15], Eimear E. Kenny[16,17,18], Barbara A. Koenig[19], Daofeng Li[1,2,3], Tobias Marschall[20],

3. Pangenome & pangenomics

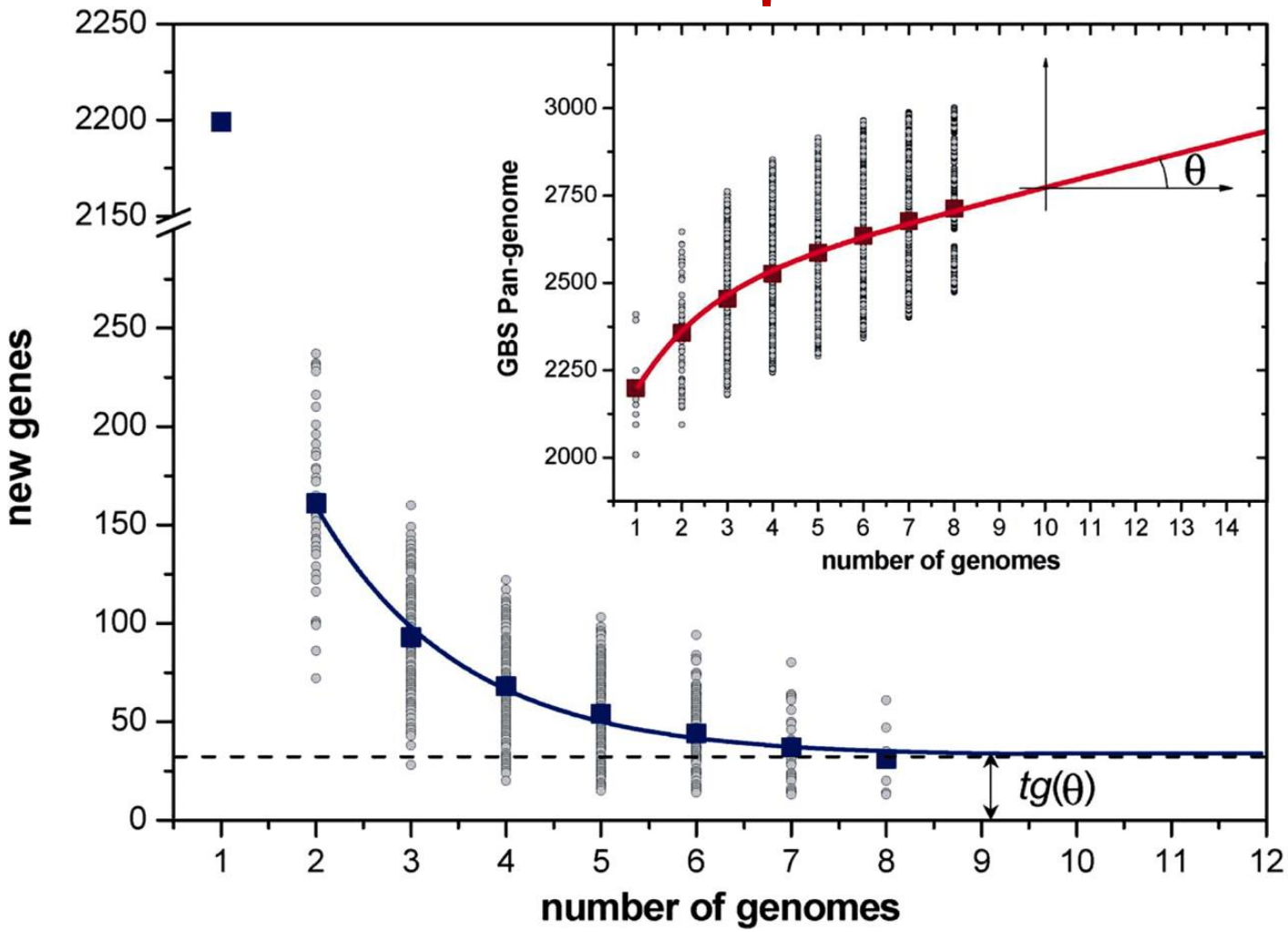
A pan-genome is the total genetic information of a population

- Individual genome
- Pan-genome*
  - Core genome c
  - Dispensable genome d d d
  - Unique genome u u u



* Tettelin, Hervé, et al. *PNAS*, 2005

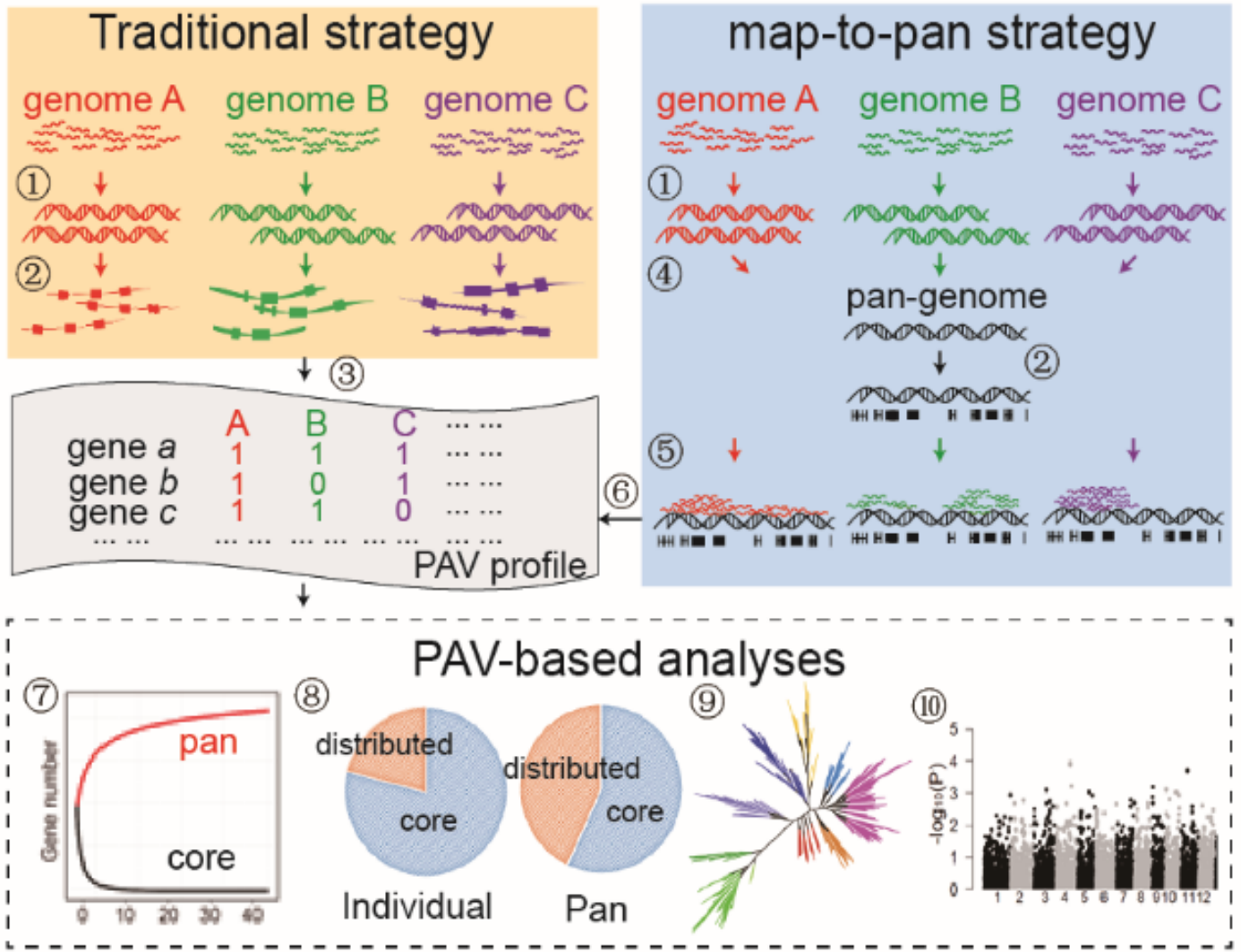
Open pan-genome of GBS：infinite number of genes!



Hervé Tettelin et al. *PNAS* 2005;102:13950-13955

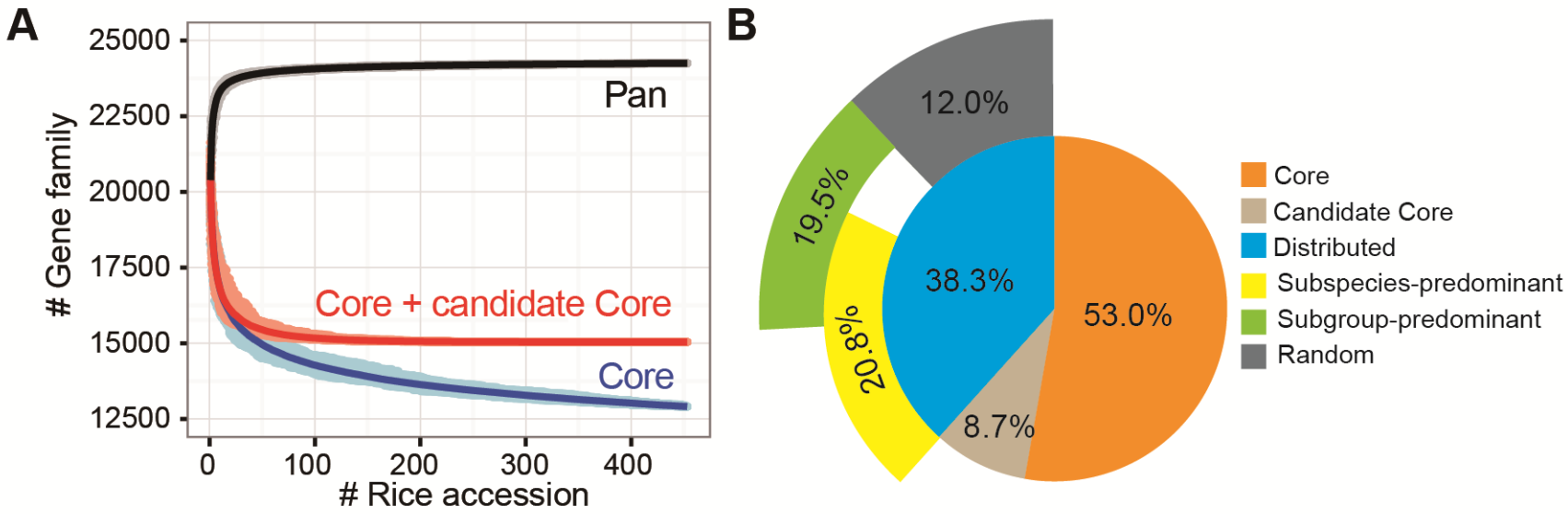
PNAS

EUPAN: Eukaryote pan-genome construction and analysis



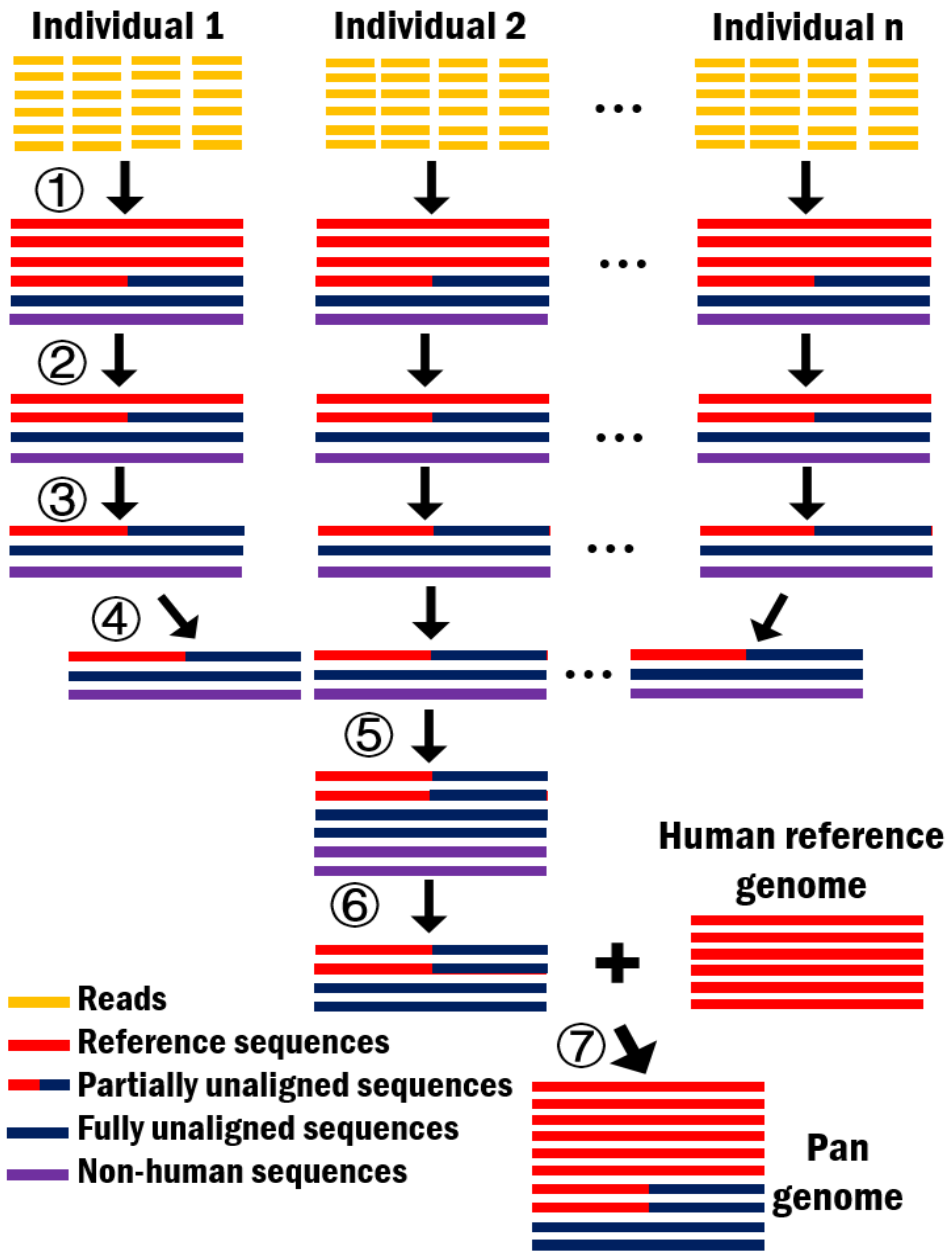
Hu, Z. et al. *Bioinformatics*, 23(15):2408-2409

3000 rice genomes→ rice pan-genome



- Rice pan-genome
  - 70% larger（with >12,000 new genes) than the reference genome
  - Important genes can be distributed
  - Greatly increased the genetic resource for rice breeding
- How about human pan-genome?

Wang, W. et al. *Nature*, 2018, 557:43-49

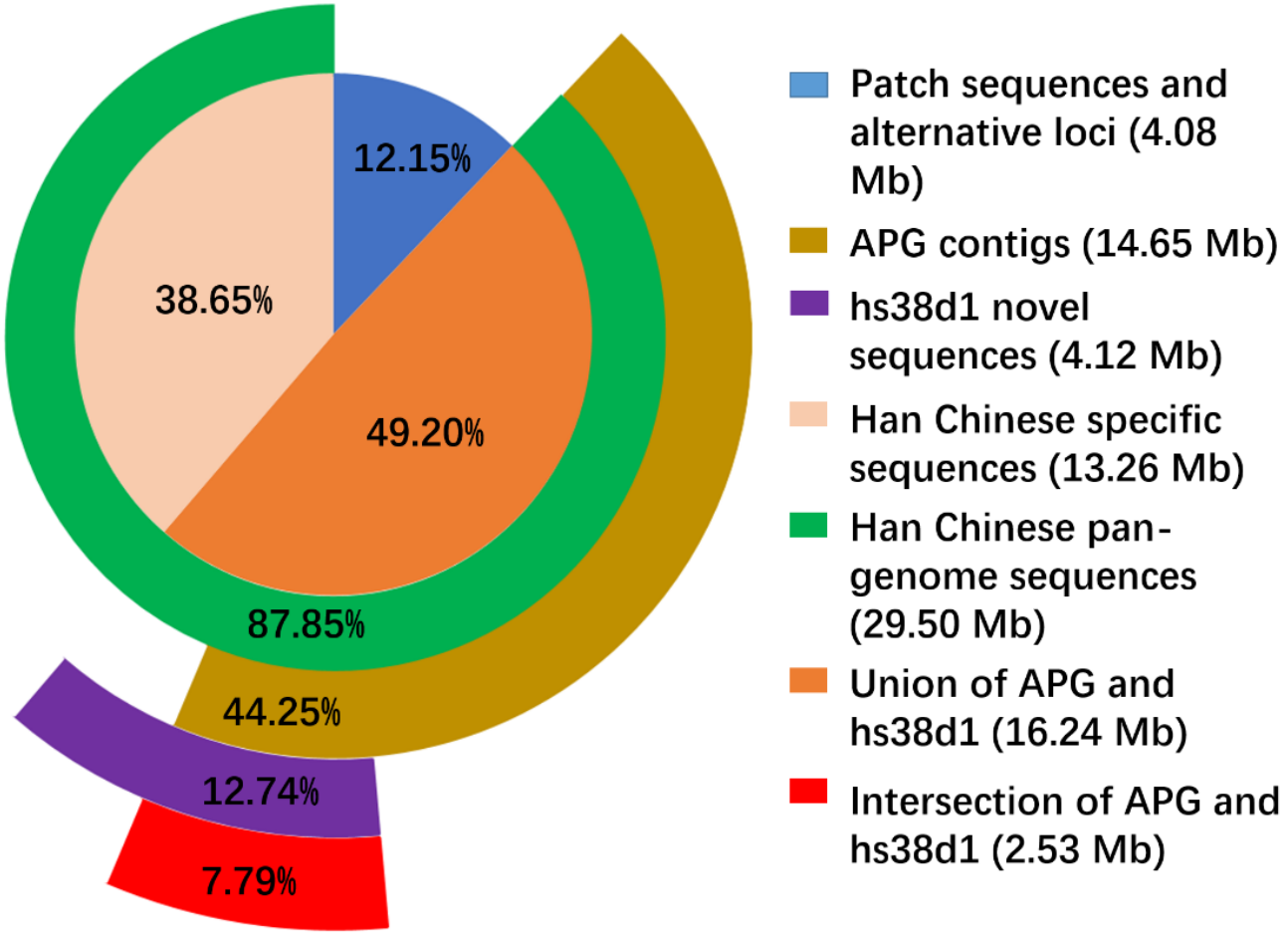


HUPAN(EUPAN2.0) for the human pan-genome analysis

① Genome assembly, GSA
② Redundancy removing
③ Extract new sequences
④ Merge
⑤ Redundancy removing
⑥ Contamination removing
⑦ Pangenome construction

Duan Z., et al. *Genome Biology,* 2019, 20:149

Comparison of Han Chinese pangenome and other human genome/pan-genomes (275 samples, 33.58Mb novel sequences)



Duan, Z. et al., ***Genome Biology***, 2019, 20:149

Comparison of pan-genome analysis tools: EUPAN1.0-3.0

Method	EUPAN	HUPAN (EUPAN2.0)	EUPAN3.0
Year	2017	2019	2022
Journal	*Bioinformatics*	*Genome Biology*	*Genome Research*
Sequencing platform	NGS	NGS	TGS
Speed	Fast	fast	Slow
Memory requirement	High（1.5TB）	low（100GB）	Middle

4. Pangenomic analysis for Chinese gastric cancer

nature communications

Article

https://doi.org/10.1038/s41467-022-33073-7

Pangenomic analysis of Chinese gastric cancer

Received: 6 March 2022

Accepted: 31 August 2022

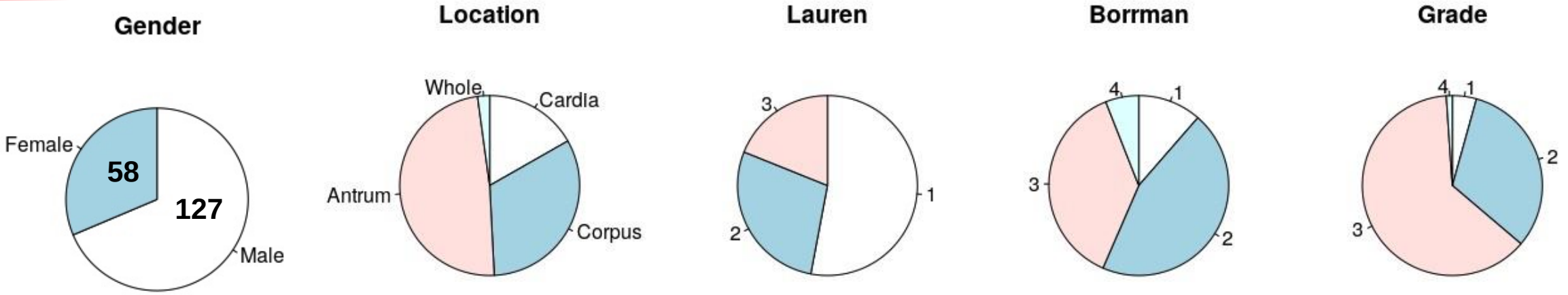
Published online: 15 September 2022

Check for updates

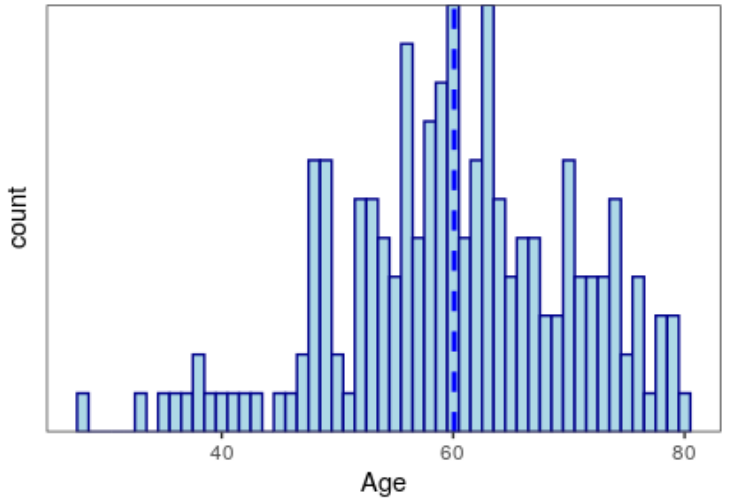
Yingyan Yu[1,10] ✉, Zhen Zhang[2,10], Xiaorui Dong[3,10], Ruixin Yang[1,10], Zhongqu Duan[3,4,10], Zhen Xiang[1], Jun Li[1], Guichao Li[2], Fazhe Yan[3], Hongzhang Xue[3], Du Jiao[3], Jinyuan Lu[3], Huimin Lu[3], Wenmin Zhang[3], Yangzhen Wei[3], Shiyu Fan[3], Jing Li[3], Jingya Jia[3], Jun Zhang[5], Jun Ji[1], Pixu Liu[6], Hui Lu[3,4], Hongyu Zhao[4], Saijuan Chen[7], Chaochun Wei[3,4] ✉, Hongzhuan Chen[8,9] ✉ & Zhenggang Zhu[1] ✉

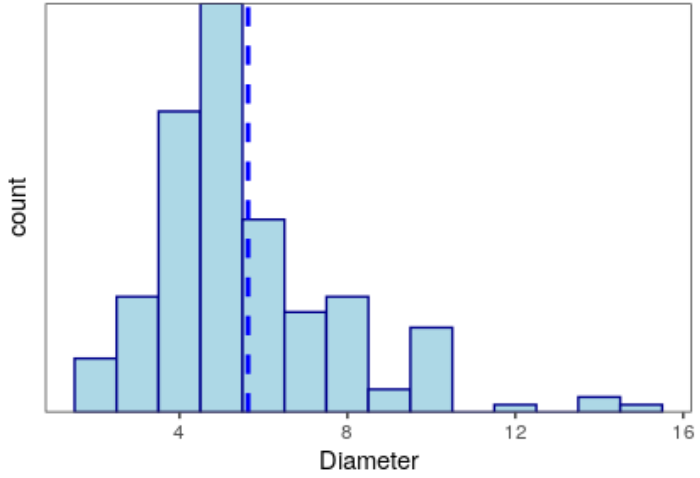
Pangenomic study might improve the completeness of human reference genome (GRCh38) and promote precision medicine. Here, we use an automated pipeline of human pangenomic analysis to build gastric cancer pangenome for 185 paired deep sequencing data (370 samples), and characterize the gene presence-absence variations (PAVs) at whole genome level. Genes *ACOT1*, *GSTM1*, *SIGLEC14* and *UGT2B17* are identified as highly absent genes in gastric cancer population. A set of genes from unaligned sequences with GRCh38 are predicted. We successfully locate one of predicted genes *GC0643* on chromosome 9q34.2. Overexpression of *GC0643* significantly inhibits cell growth, cell migration and invasion, cell cycle progression, and induces cell apoptosis in cancer cells. The tumor suppressor functions can be reversed by sh*GC0643* knockdown. The *GC0643* is approved by NCBI database (GenBank: MW194843.1). Collectively, the robust pan-genome strategy provides a deeper understanding of the gene PAVs in the human cancer genome.

185 pairs of samples(cancer and normal tissues)



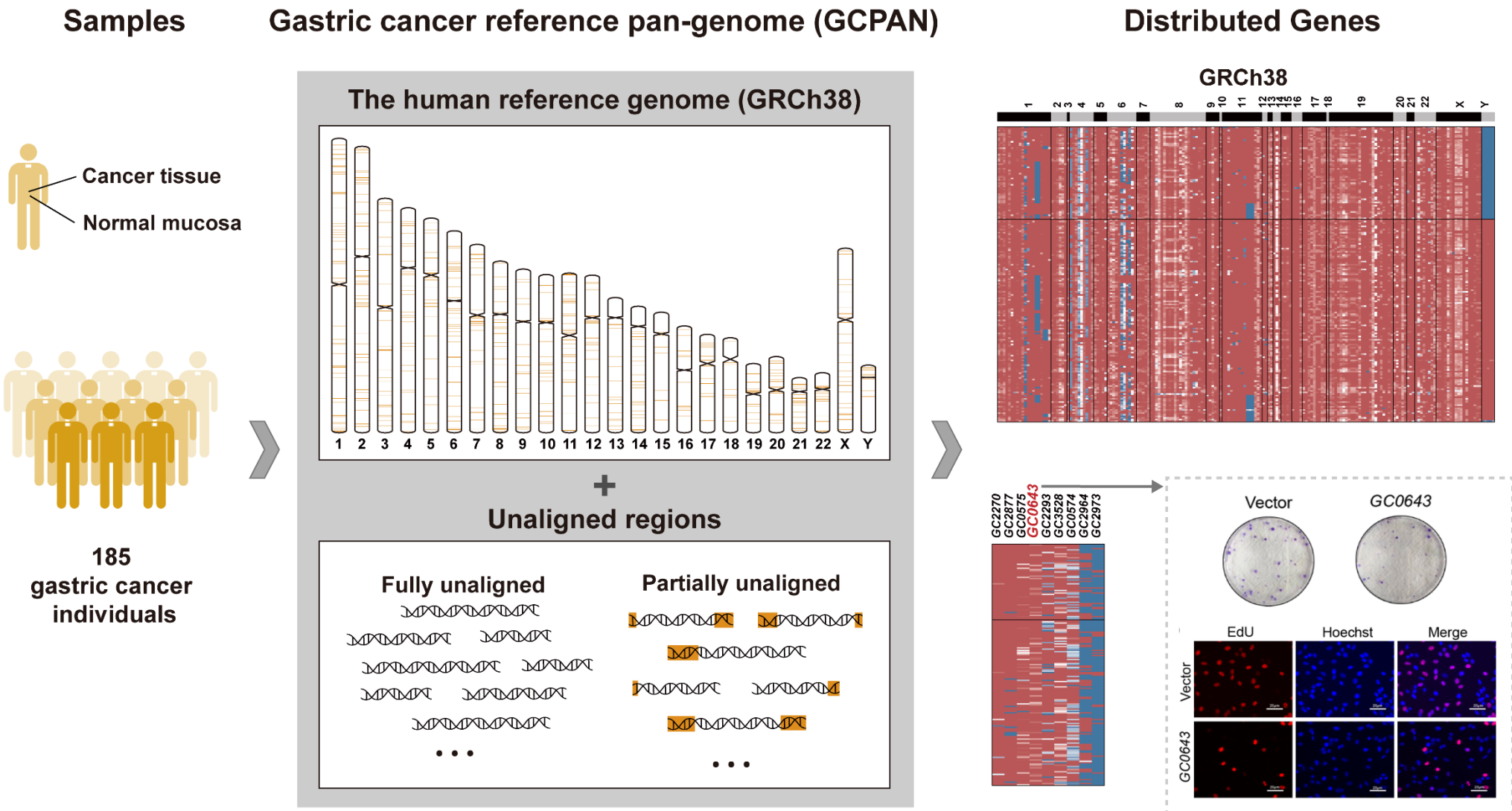
WGS	Reads depth	Reads length	Contig N50
Normal	30X	150bp	8,042bp
Tumor	60X	150bp	7,889bp





Sequence data 50Tb in total

10 phenotypes: age, gendre, typing (18 types, Lauren, Borrman), location, size, stage, microbe infection (HP and EBV)

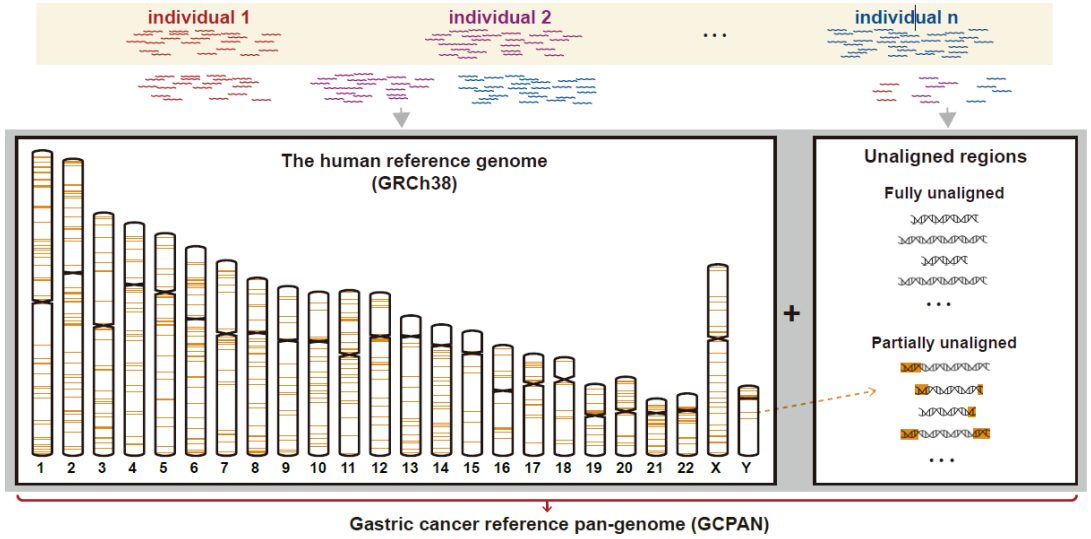


Pangenomic analysis of Chinese gastric cancer, Yu et al., *Nature Communications*, 2022

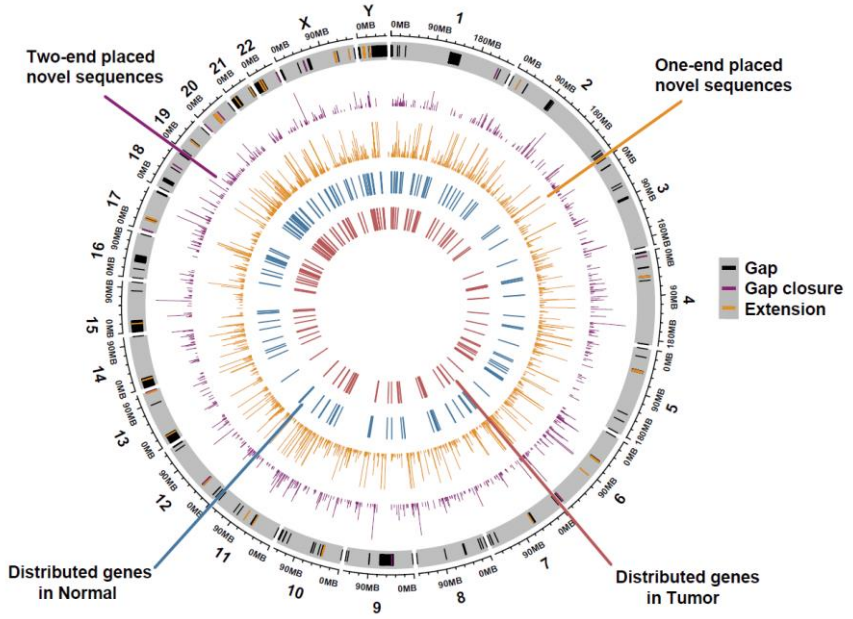
Construction of gastric cancer pangenome (GCPAN)

a. GCPAN include:

- ❖ The reference genome GRCh38
- ❖ unaligned（80.88Mb, >500bp）
  - ❖ Partially unaligned
  - ❖ Fully unaligned



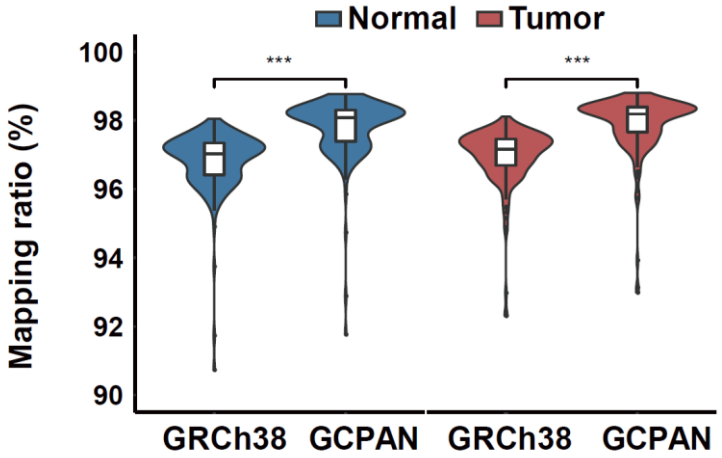
b. Distribution of partially unaligned regions and gene PAVs among the human reference genome GRCh38



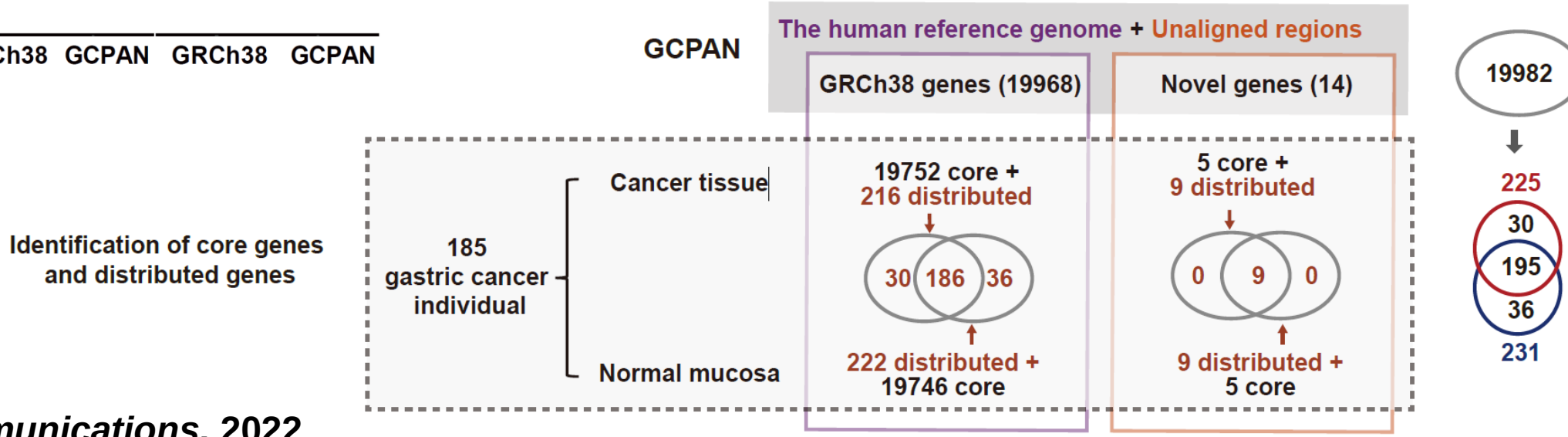
Yu et al., *Nature Communications*, 2022

Han Chinese gastric cancer pan-genome

Mapping rate is increased significantly ($p < 2.22e-16$)



Core genes (19721)
Distributed genes (261)

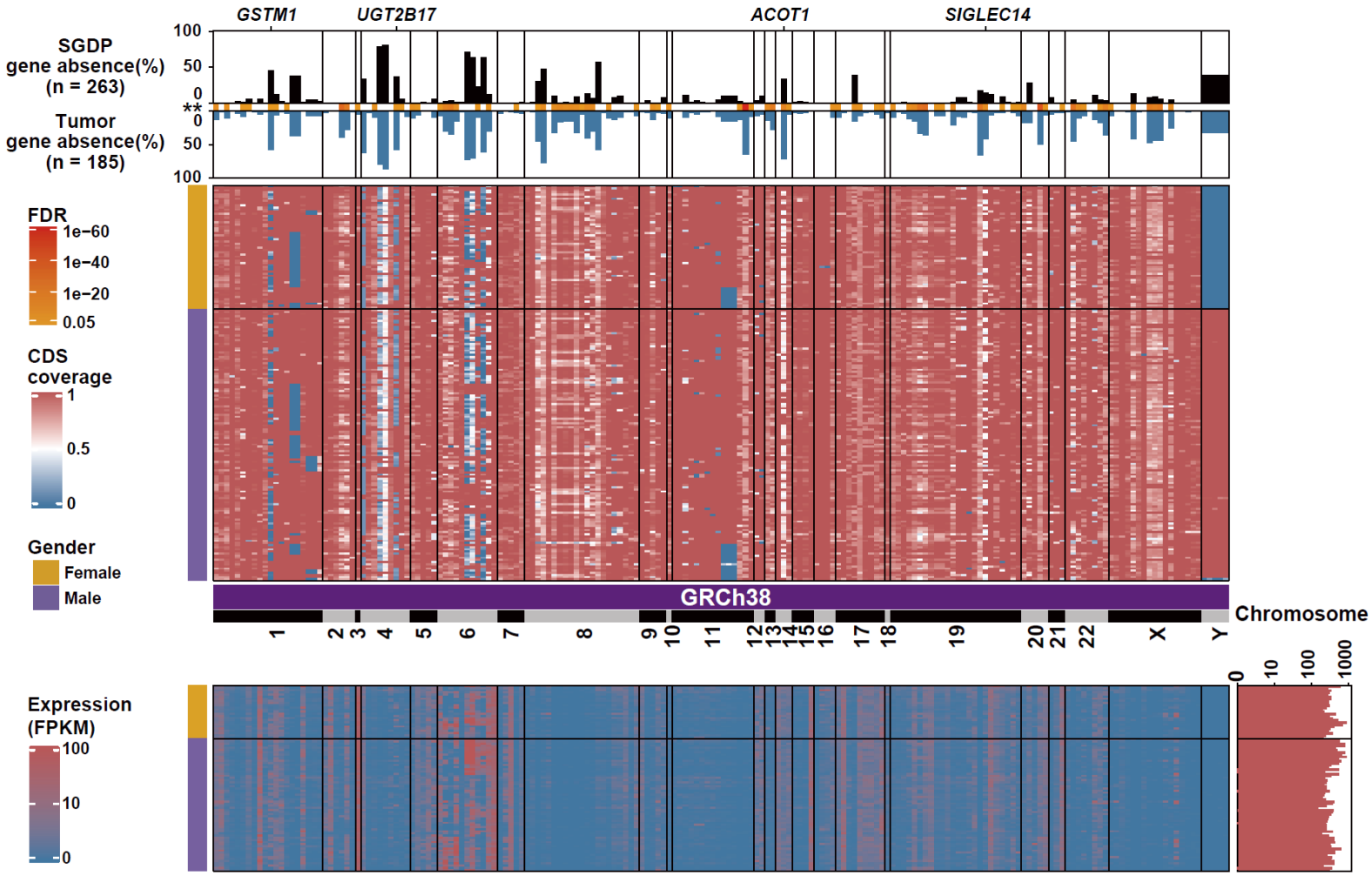


Yu et al., *Nature Communications*, 2022

Gene PAVs on the human reference genome

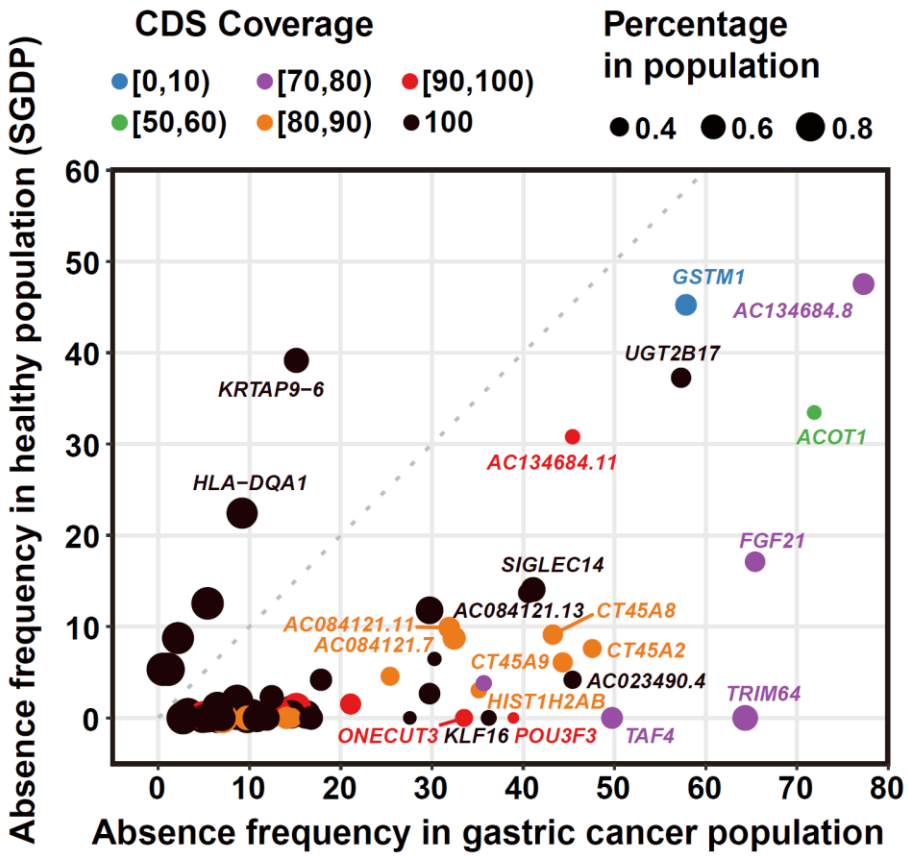
a. Distributed genes

b. Gene expression level



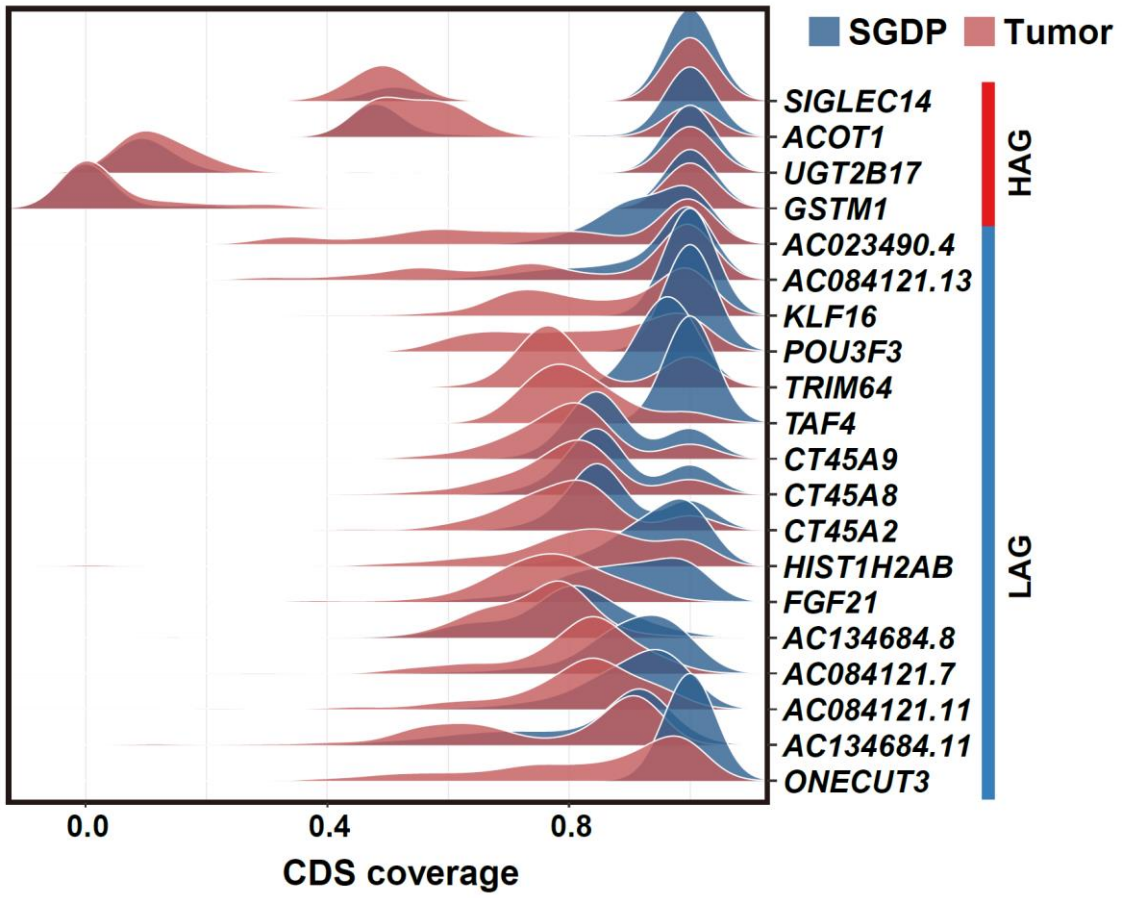
Yu et al., *Nature Communications*, 2022

Gene PAVs in the human reference genome



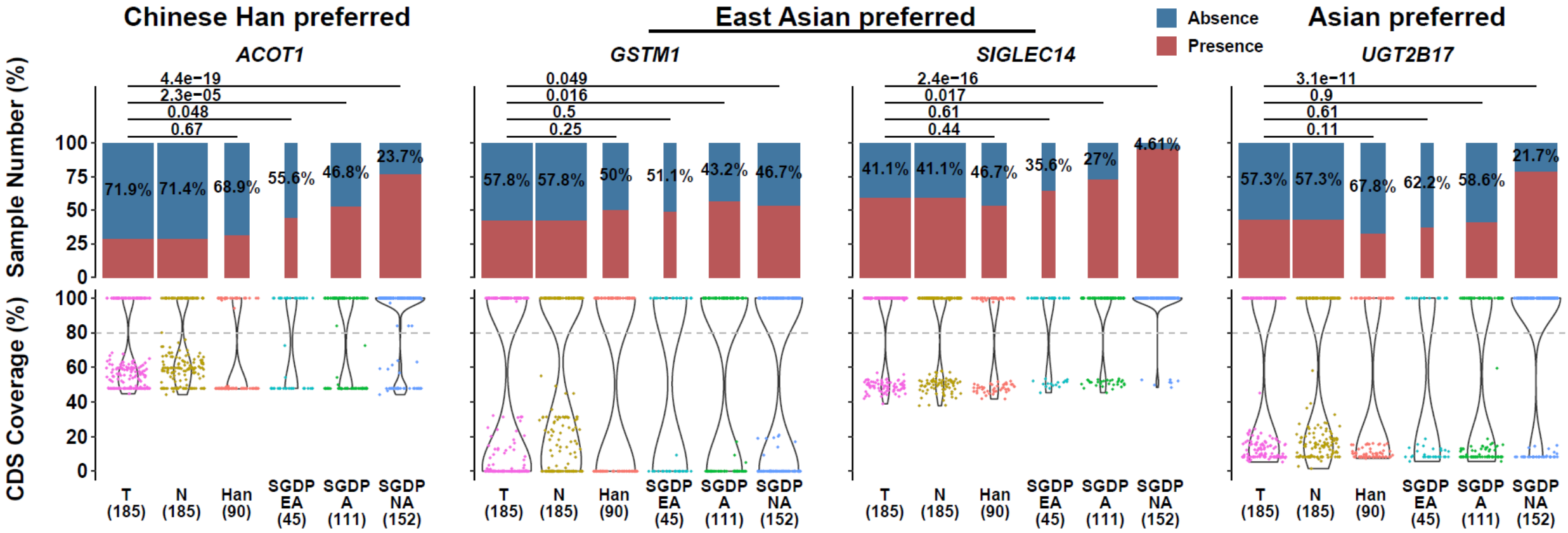
Comparison of gene absence frequency between cancer and healthy populations

The most frequently and severely absent genes



Yu et al., *Nature Communications*, 2022

Gene PAVs in the human reference genome

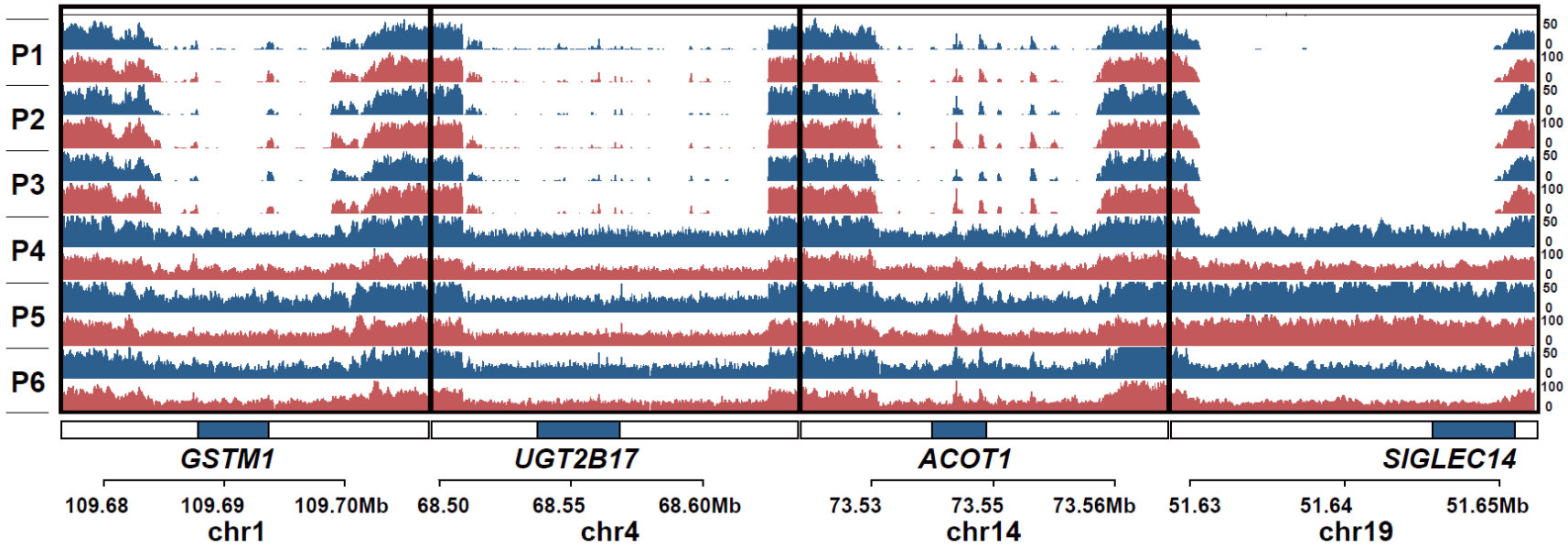


Comparison of gene PAV frequencies in different populations for 4 most severely absent genes

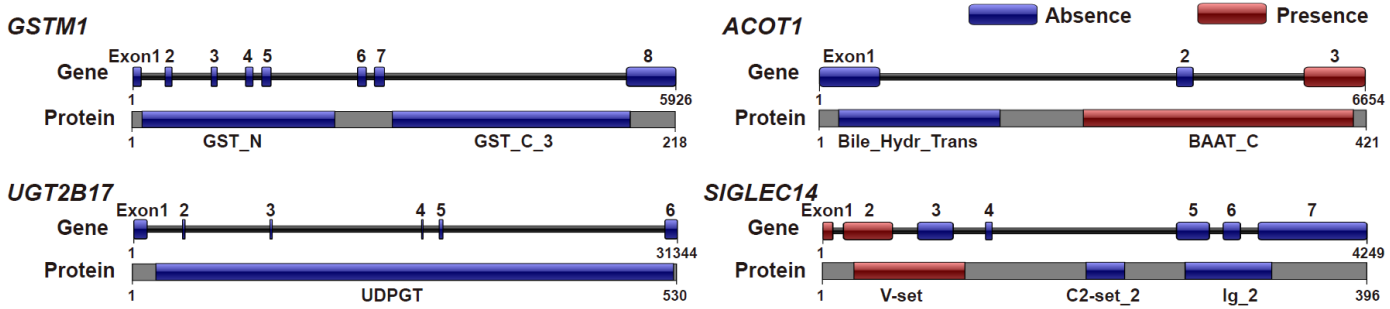
Yu et al., *Nature Communications*, 2022

Gene PAVs for 4 most absent genes

- Genomic sequence



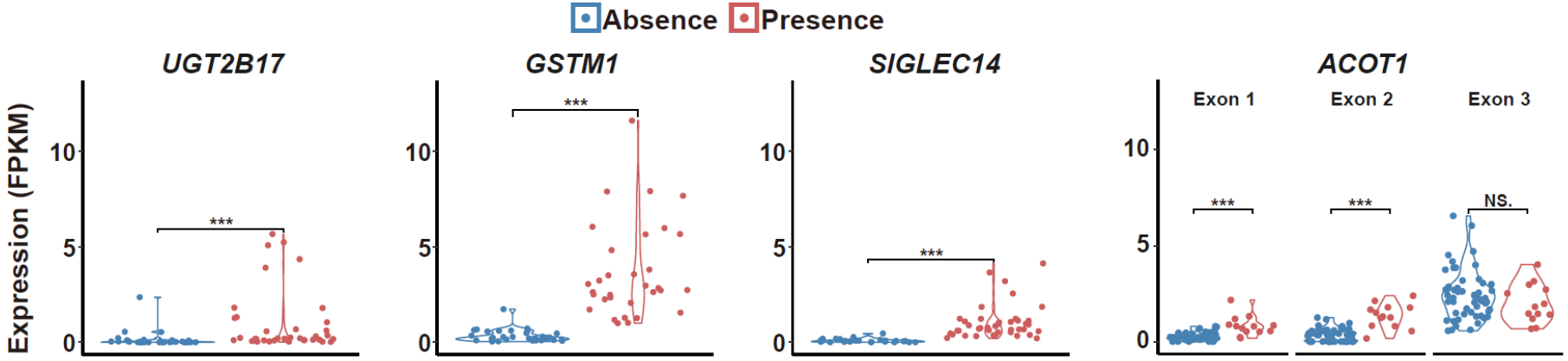
- Gen structure and CDS PAVs



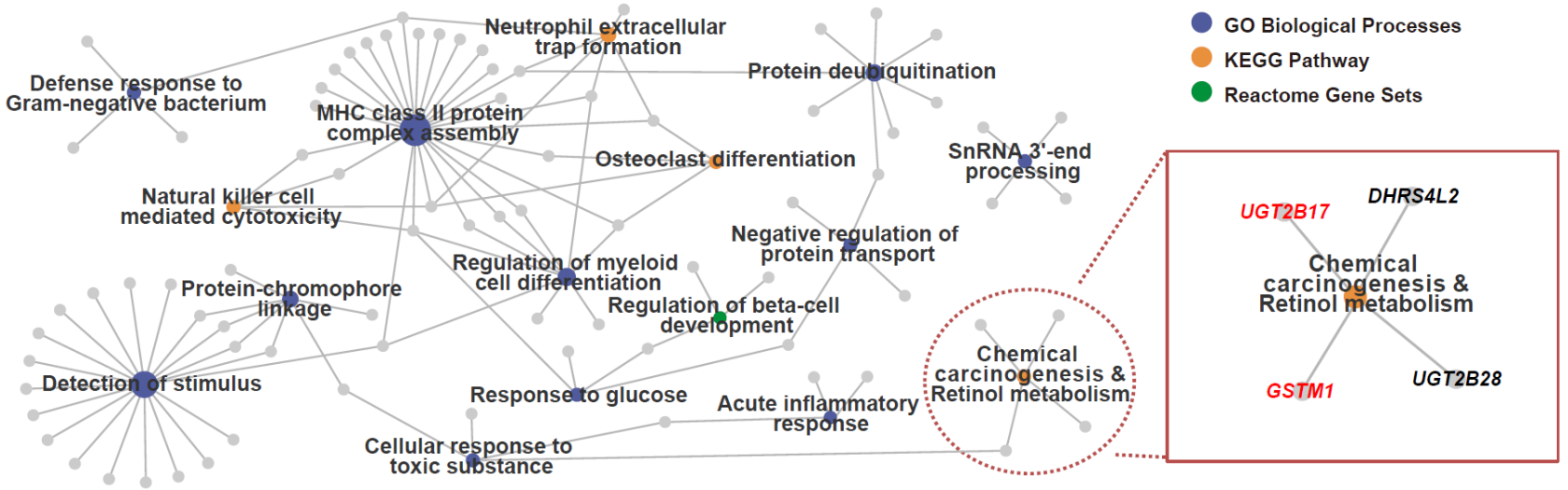
Yu et al., *Nature Communications*, 2022

The gene PAVs of the top four most absent genes

- Distribution of expression levels in the cancer samples



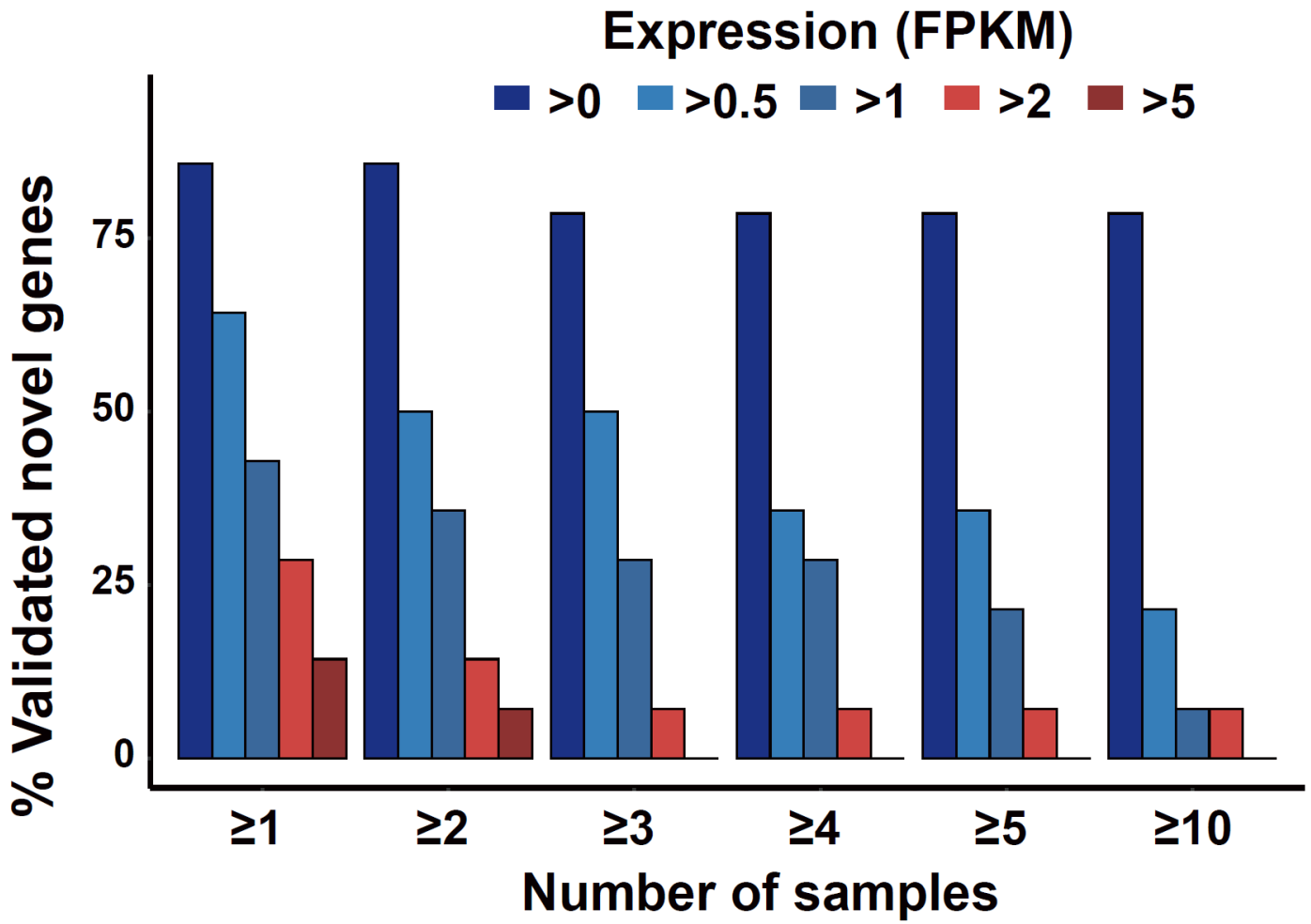
- 186 distributed genes and their metabolic pathways



Yu et al., *Nature Communications*, 2022

Genes in non-reference genomic regions

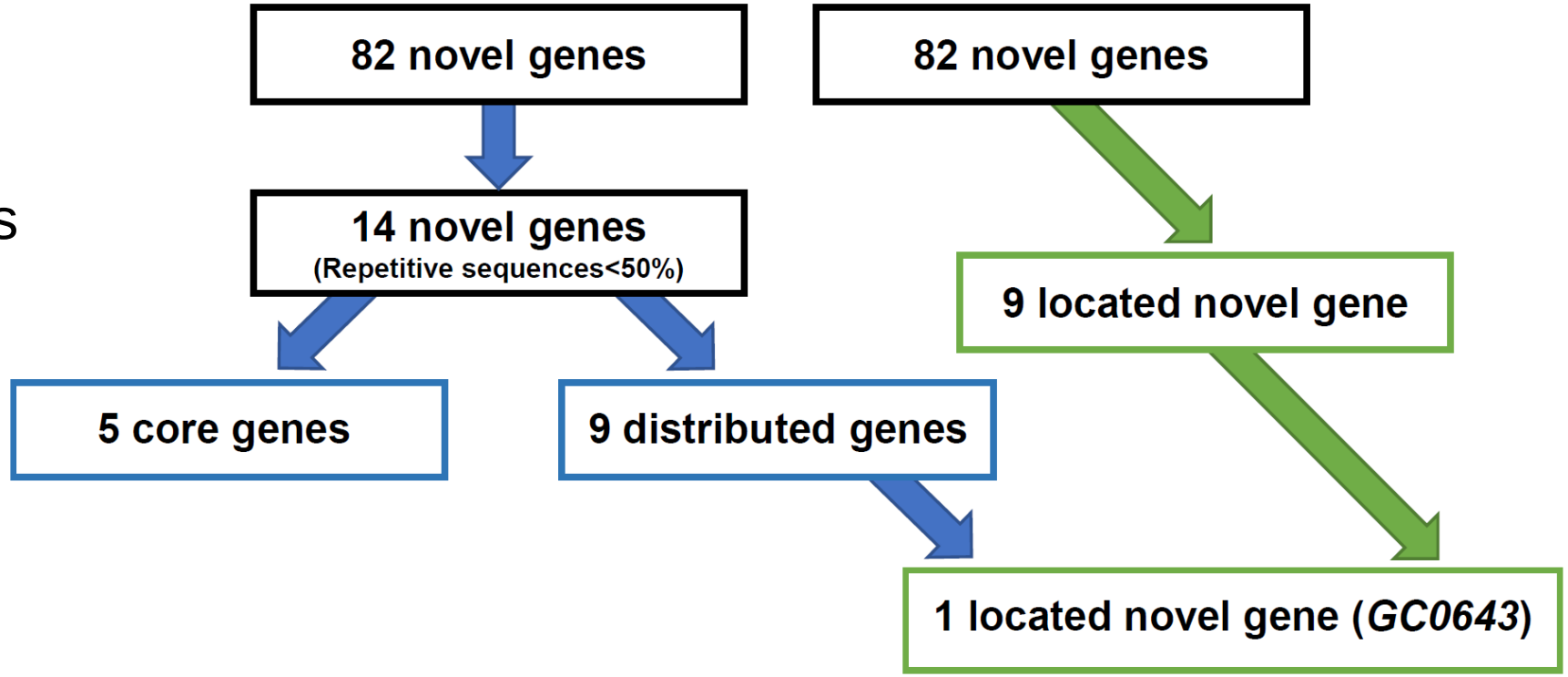
Validation at transcriptome level



Yu et al., *Nature Communications*, 2022

New genes in non-reference genomic regions

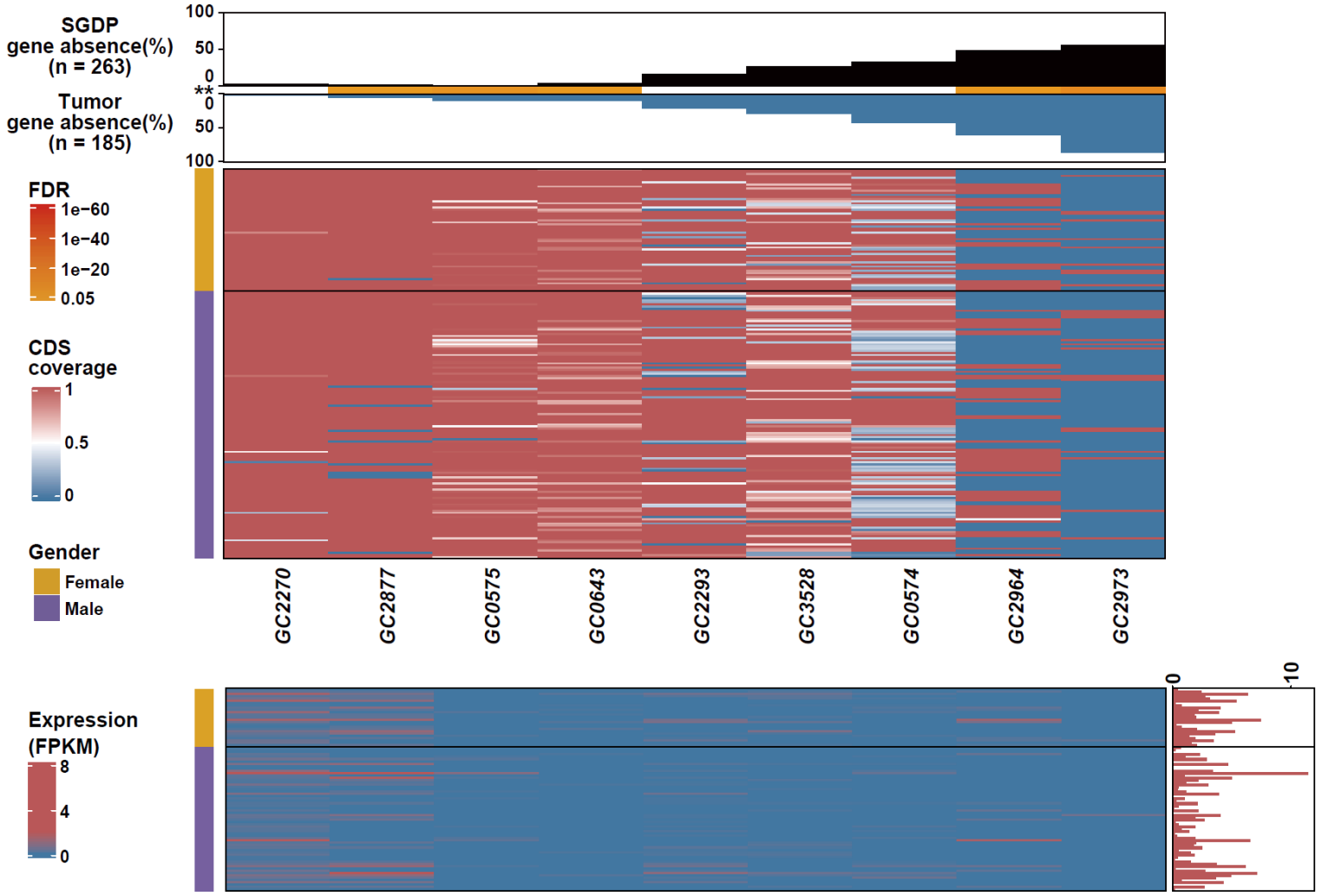
PAV analysis and their chromosomal locations



Yu et al., *Nature Communications*, 2022

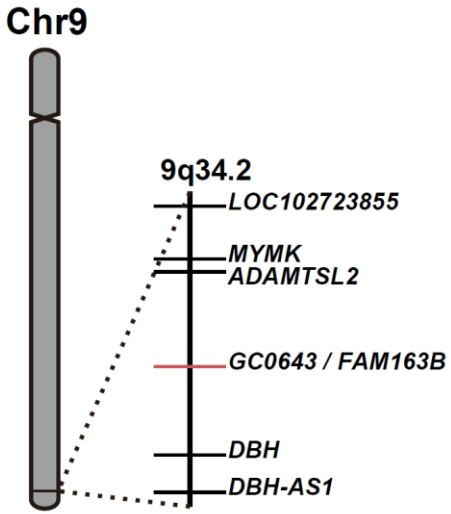
Gene PAVs in non-reference genomic regions

- Genomic level
- Transcript level

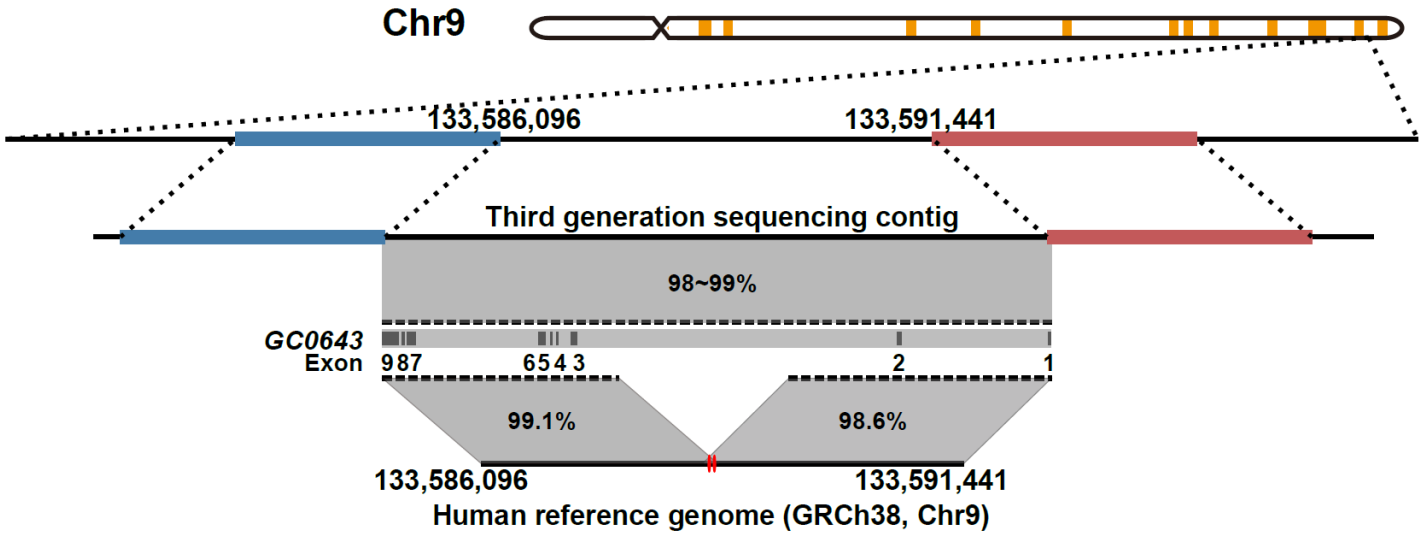


Yu et al., Nature Communications, 2022

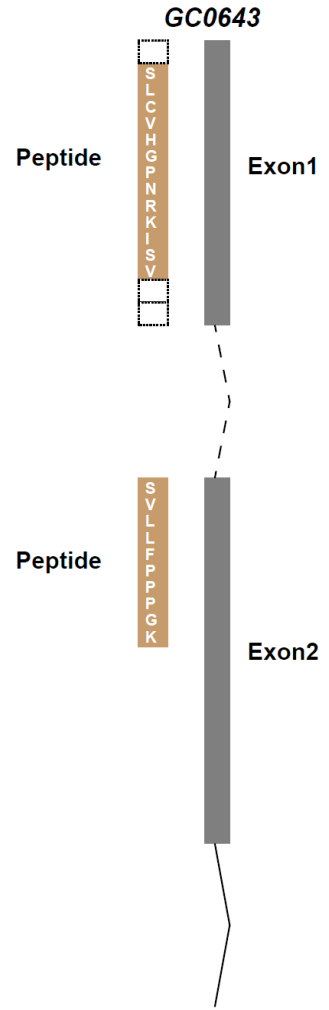
New gene GC0643



Location in chromosomes



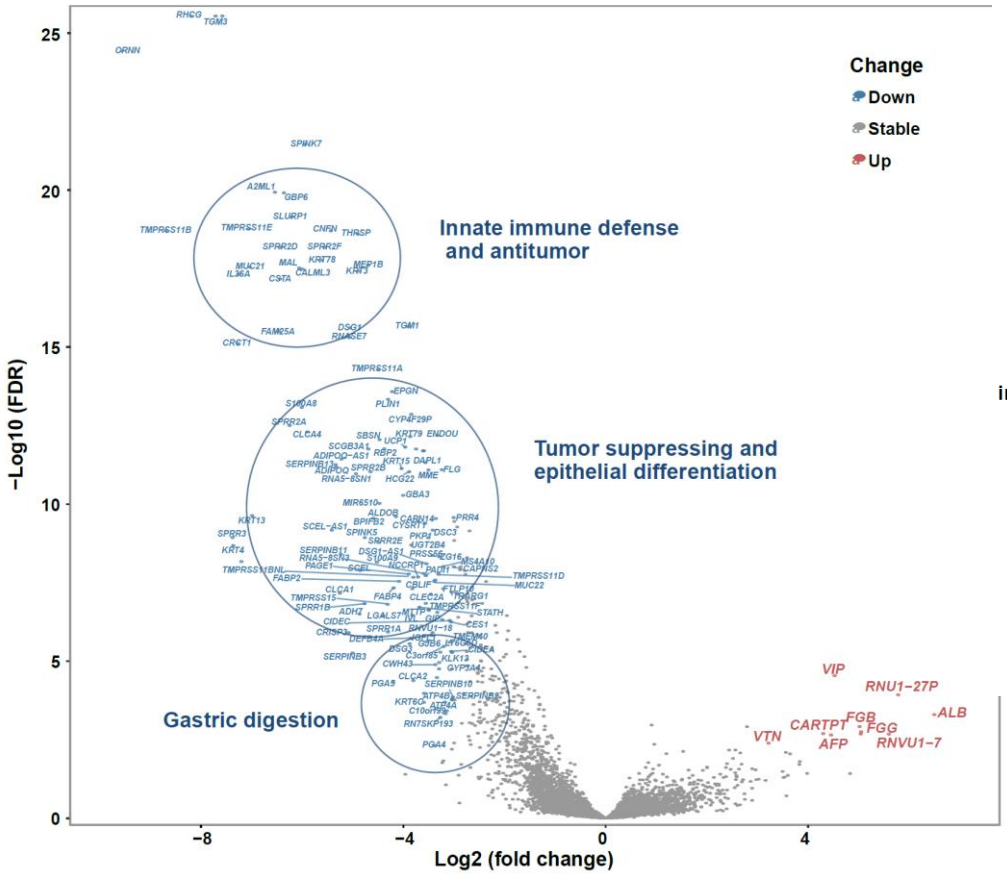
Comparison of genomic variations、gene structure and long sequencing reads



Proteomic evidence

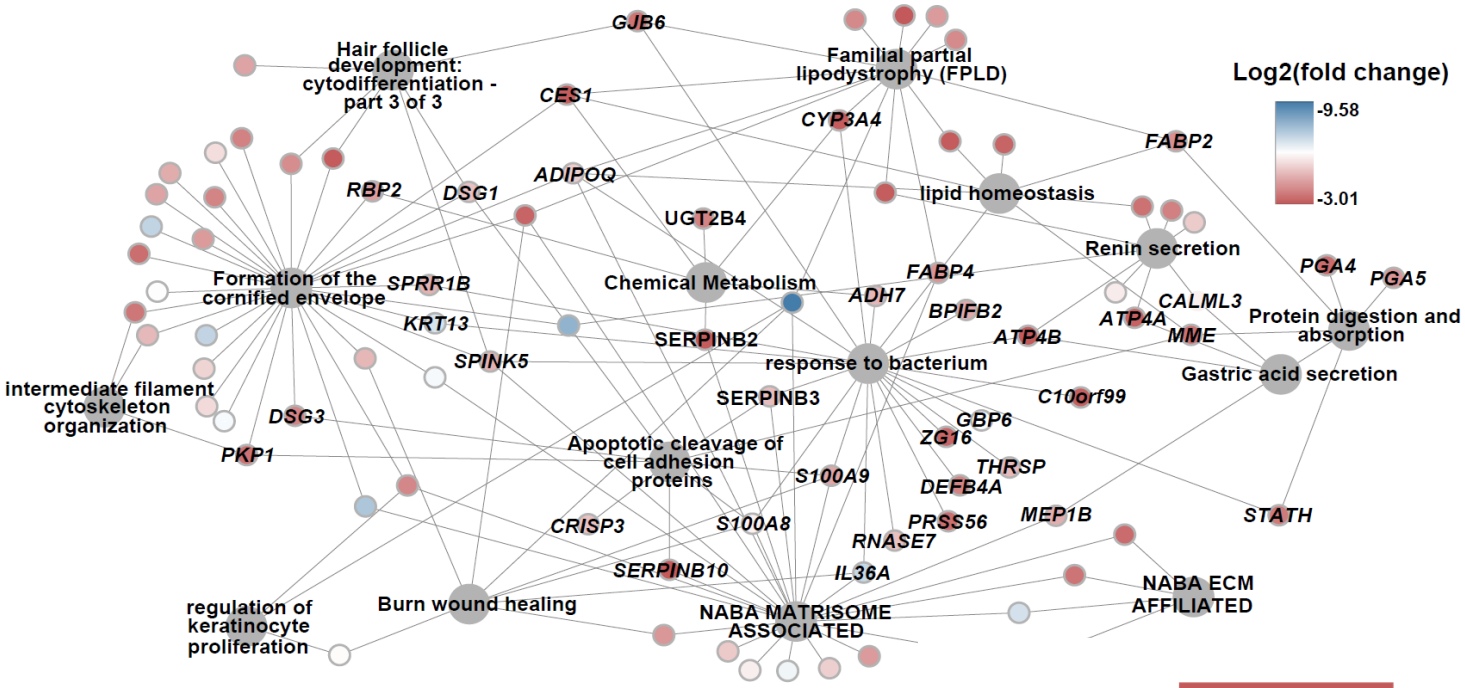
Yu et al., Nature Communications, 2022

New gene GC0643

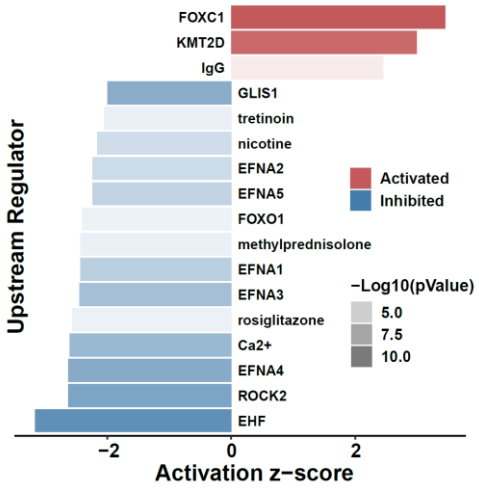


Differential gene express analysis：

GC0643 gene presence samples vs gene absence samples

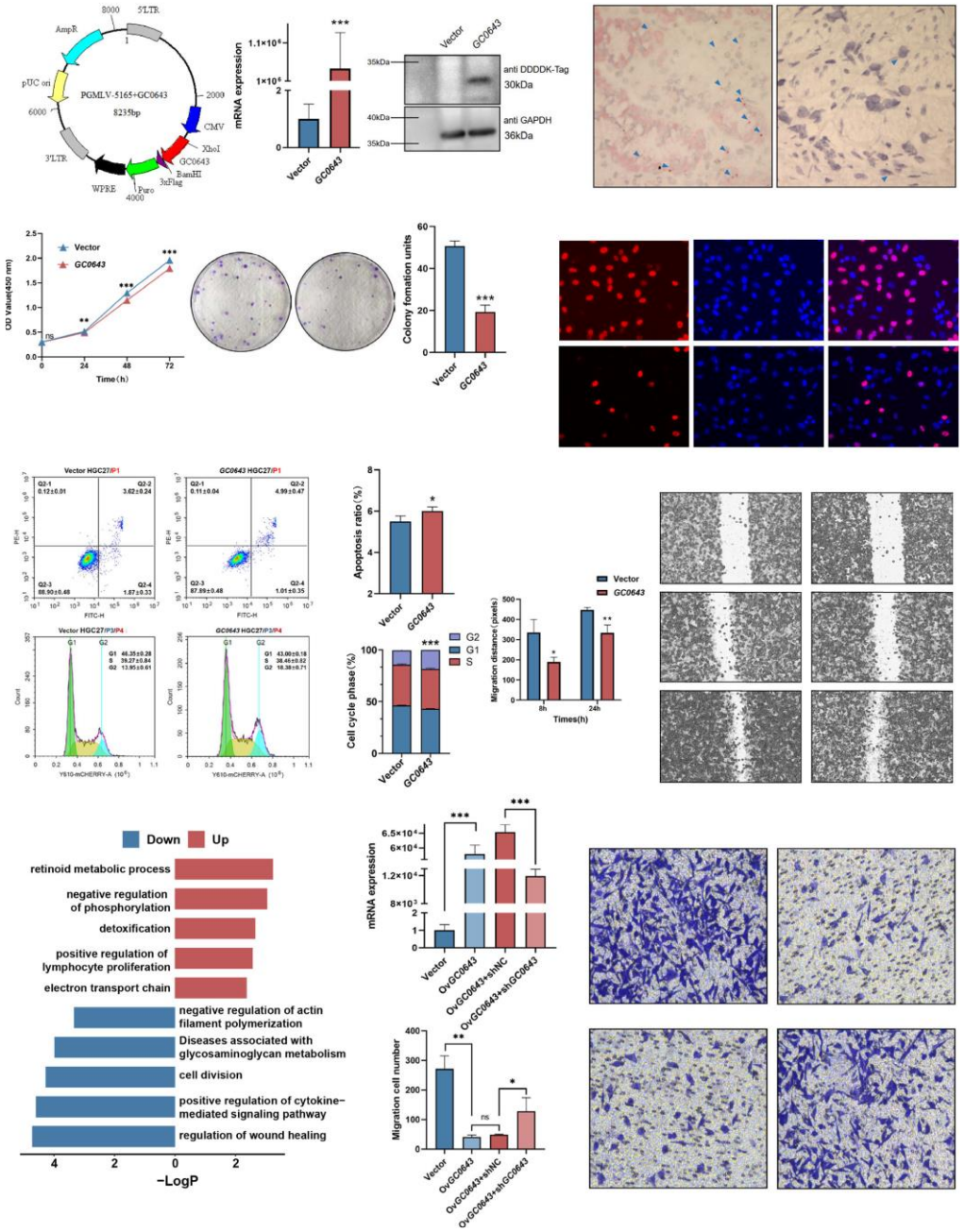


Enrichment analysis of differentially expressed genes



Yu et al., Nature Communications, 2022

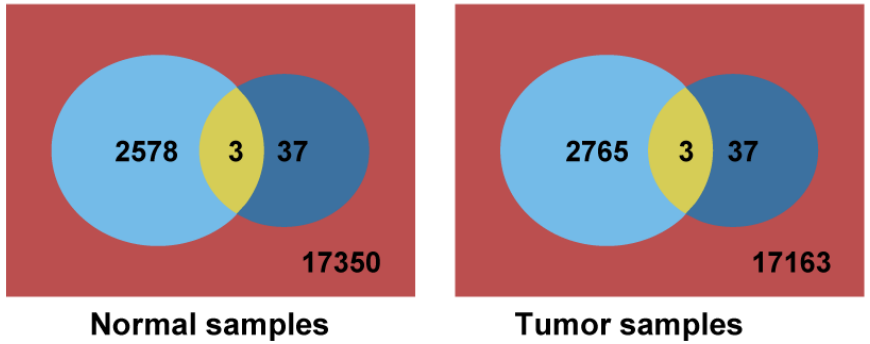
Function of GC0643：reduce cancer cell growth



Yu et al., Nature Communications, 2022

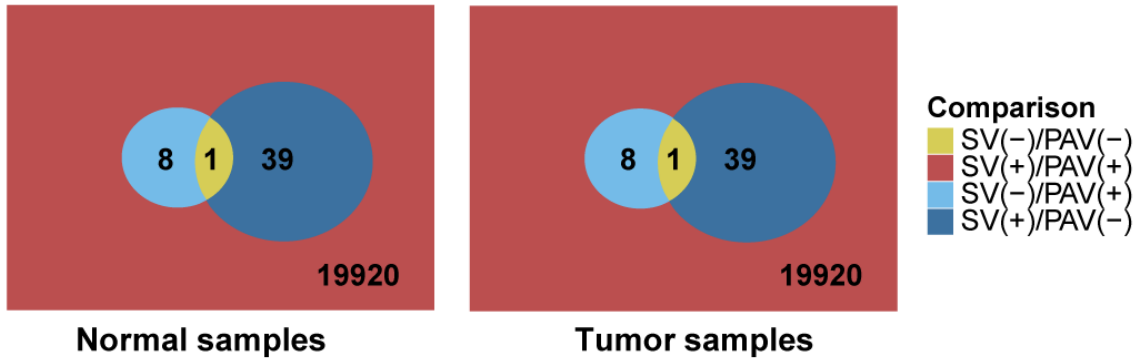
Complementary to traditional SV analysis methods

A



All DELs calculated

B



Only homozygous DELs calculated

Yu et al., Nature Communications, 2022

Data Availability

The raw sequencing data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation (GSA-Human)[https://ngdc.cncb.ac.cn/gsa-human]. The accession number for genomic sequencing data of normal gastric mucosa is HRA002344 and that for genomic and transcriptomic sequencing data of gastric cancer is HRA002333. The raw sequencing data are also available in NODE database [http://www.biosino.org/node] with the accessions OEP000301 for genomic sequencing data of normal gastric mucosa and OEP000482 for genomic and transcriptomic sequencing data for gastric cancer. The raw sequencing data are available under restricted access due to data privacy laws. Readers can get access to data by sending request to corresponding authors. Data will be available within a week once the access has been granted.

The processed data and result files are available on the website http://cgm.sjtu.edu.cn/cpan/GCPAN.html.

Data that support the findings of this study are available at http://gigadb.org/dataset/100302 for the sequencing data of 90 Han Chinese, https://www.ebi.ac.uk/ena/browser/view/PRJEB9586 for the SGDP data, https://pdc.cancer.gov/pdc/browse/filters/primary_site:Stomach with study ID PDC000214 for the proteomics data for gastric cancer from the CPTAC project and https://www.ncbi.nlm.nih.gov/sra with study IDs PRJNA301527, PRJA339722, PRJNA530217, and PRJNA551670 for the long-read sequencing data of humans to locate the non-reference genes to their corresponding chromosome positions. Source data are provided with this paper.

Summary

1. First cancer pan-genomics analysis
    - 80.88Mb non-reference sequences (at least 14 new genes)
    - 261 dispensable genes (~1%)
        - 195 shared by tumor and mucosa
            - 186 located in the human reference genome, 9 new
2. Four gene absence variations are more frequent in Chinese population
    - *GSTM1*, UTG2B17*, ACOT1*, SIGLEC14* (*cancer related)
3. At least 14 non-reference genes
    - GC0643 inhibits cancer cell growth
4. Complementary to traditional disease genome analysis methods

This study can partially explain the high rate of gastric cancer in Chinese.

5. Discussion & future plan

ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

The Need for a Human Pangenome Reference Sequence

TECHNICAL REPORT

https://doi.org/10.1038/s41588-022-01043-w

nature genetics

Check for updates

OPEN

Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes

Jana Ebler[1], Peter Ebert[1], Wayne I
Torsten Houwaart[6], Yafei Mao[5], J
Alexander T. Dilthey[6,8,9] and Tobias M

**Typical genotyping workflows map reads to a
ments introduces reference biases and comes
the ability to characterize repetitive genomic**

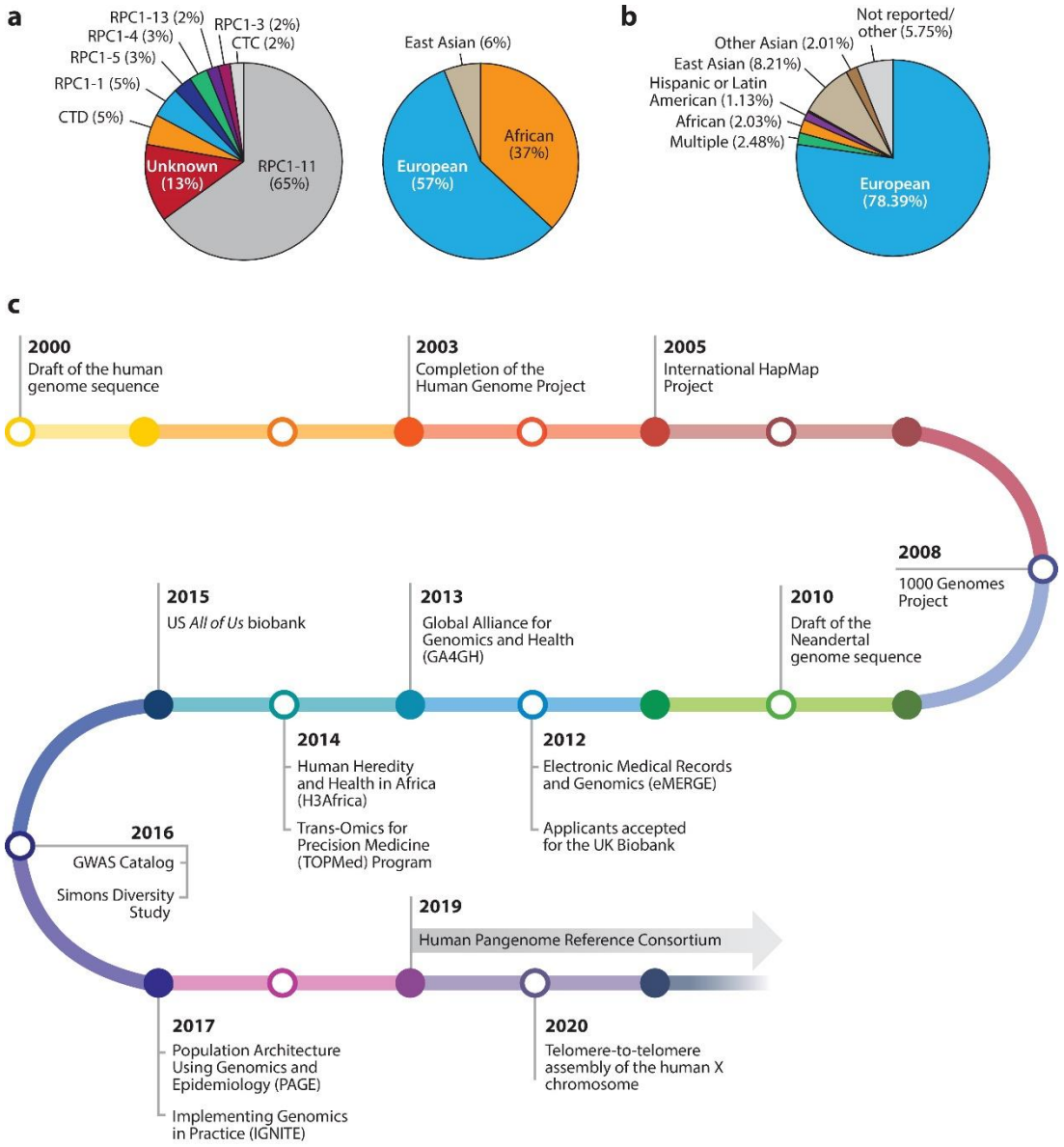
Annu. Rev. Genom. Hum. Genet. 2021. 22:
First published as a Review in Advance on April 30, 2021
The *Annual Review of Genomics and Human*

Perspective

The Human Pangenome Project: a global resource to map genomic diversity

https://doi.org/10.1038/s41586-022-04601-8

Ting Wang[1,2,3]✉, Lucinda Antonacci-Fulton[3], Kerstin Howe[4], Heather A. Lawson[1],



Miga KH, Wang T. 2021
Annu. Rev. Genom. Hum. Genet. 22:81–102

Historic progress over the last 20 years has enabled the launch of the human pangenome reference initiative.

High quality reference pangenome is needed !

Annual Reviews, 2021

The non-reference genome in the human pan-genome may >>10%

The size of rice pan-genome：

NGS：380Mb + 268Mb

TGS：380Mb + 604Mb

Research

Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes

Fan Zhang,[1,2,7] Hongzhang Xue,[3,7] Xiaorui Dong,[3] Min Li,[2] Xiaoming Zheng,[1] Zhikang Li,[1,2] Jianlong Xu,[1,4] Wensheng Wang,[1,2,5] and Chaochun Wei[3,6]

[1]*Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China;* [2]*College of Agronomy, Anhui Agricultural University, Hefei 230036, China;* [3]*Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China;* [4]*Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese*

Acknowledgement

- Zhongqu Duan
- Xiaorui Dong
- Zhiqiang Hu
- Chen Sun
- Fazhe Yan
- Hongzhang Xue
- Du Jiao
- Huimin Lu
- Jinyuan Lu
- Wenmin Zhang
- Yangzhen Wei
- Shiyu Fan
- Jingya Jia
- Jing Li (SJTU)
- Hui Lv (SJTU)
- Hongyu Zhao (Yale)
- Yingyan Yu (Ruijin Hospital)
- Zhen Zhang (Shanghai Cancer Center)
- Zhenggang Zhu (Ruijin Hospital)
- Hongzhuan Chen (SJTU)

- Funding
  - NSFS
  - NSFC
  - 973
  - 863