

The Korean Genome Project Using NGS and Genome Engine Informatics Pipeline

SangHoon Song and Jong Bhak

TheragenEtex

Theraen Bio Institute(TBI), suwon 443-270, Korea

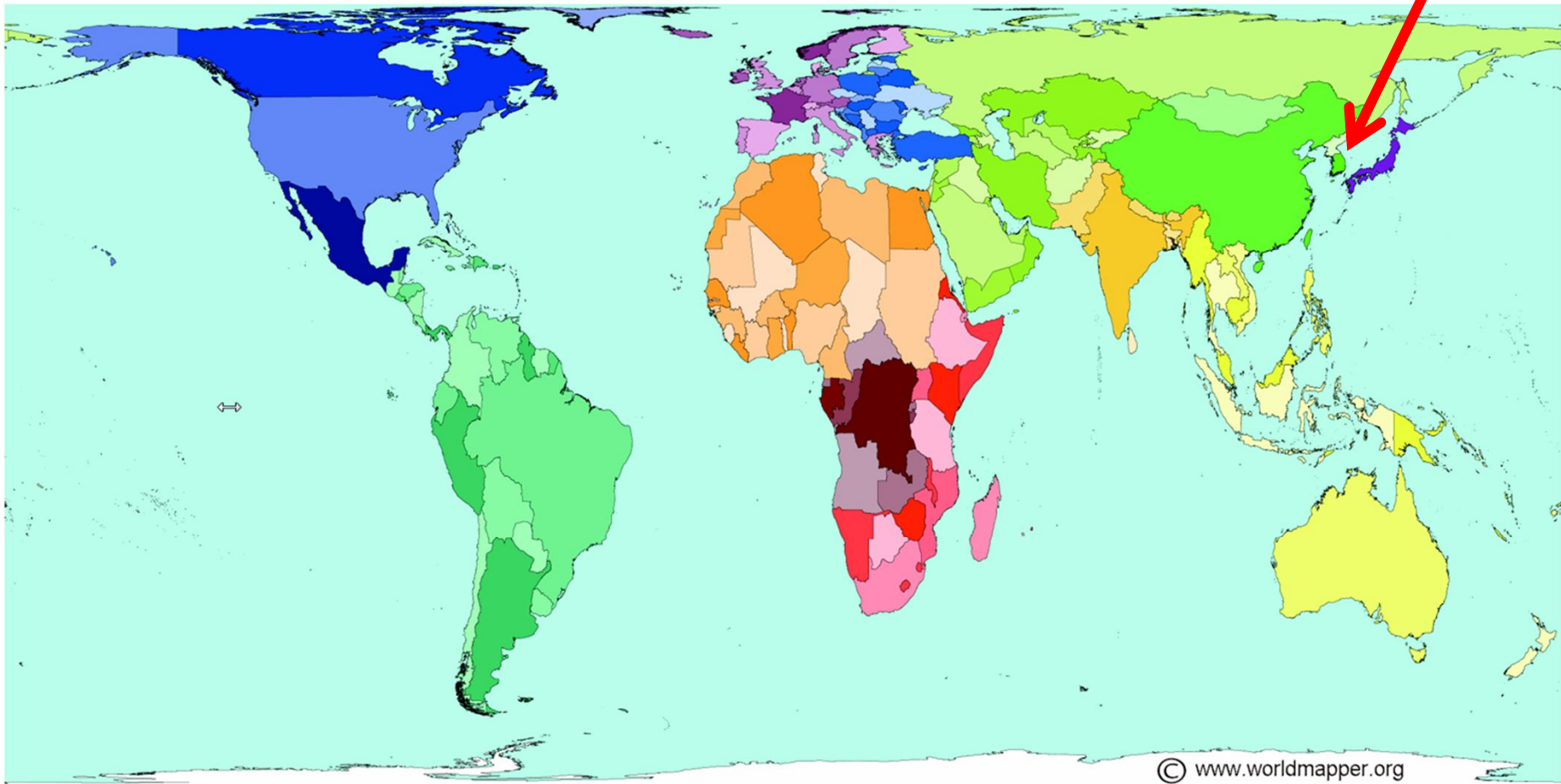


Acknowledgement

- Scientists
- Tax payers
- BioIT Asia
- Ming Guo
- Kevin Davies
- TheragenEtex (<http://therabio.org>)
- Industrial Strategic technology development program, 10040231, "Bioinformatics platform development for next generation bioinformation analysis" by the Ministry of Knowledge Economy (MKE, Korea).

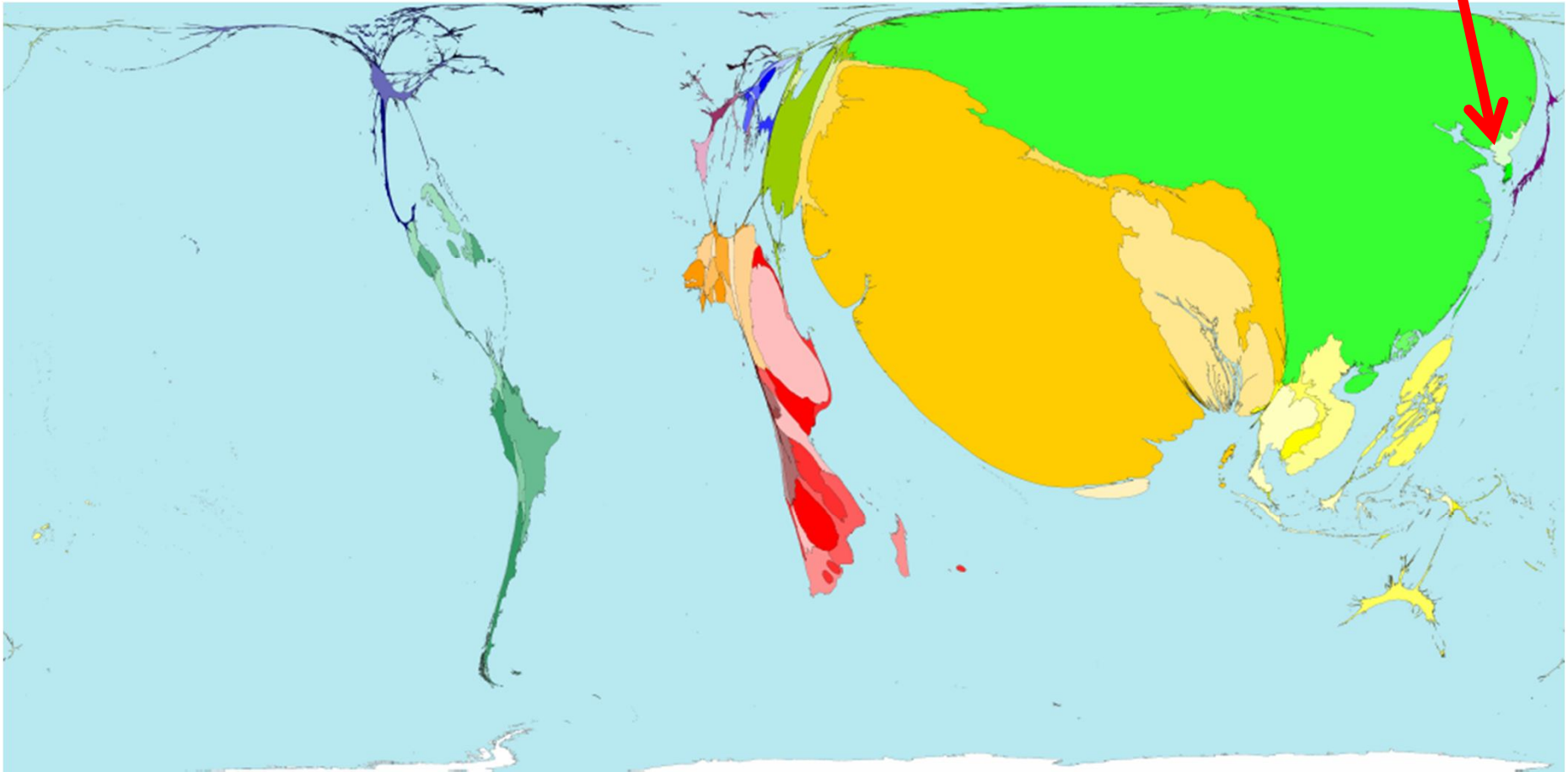


World Land Area Map



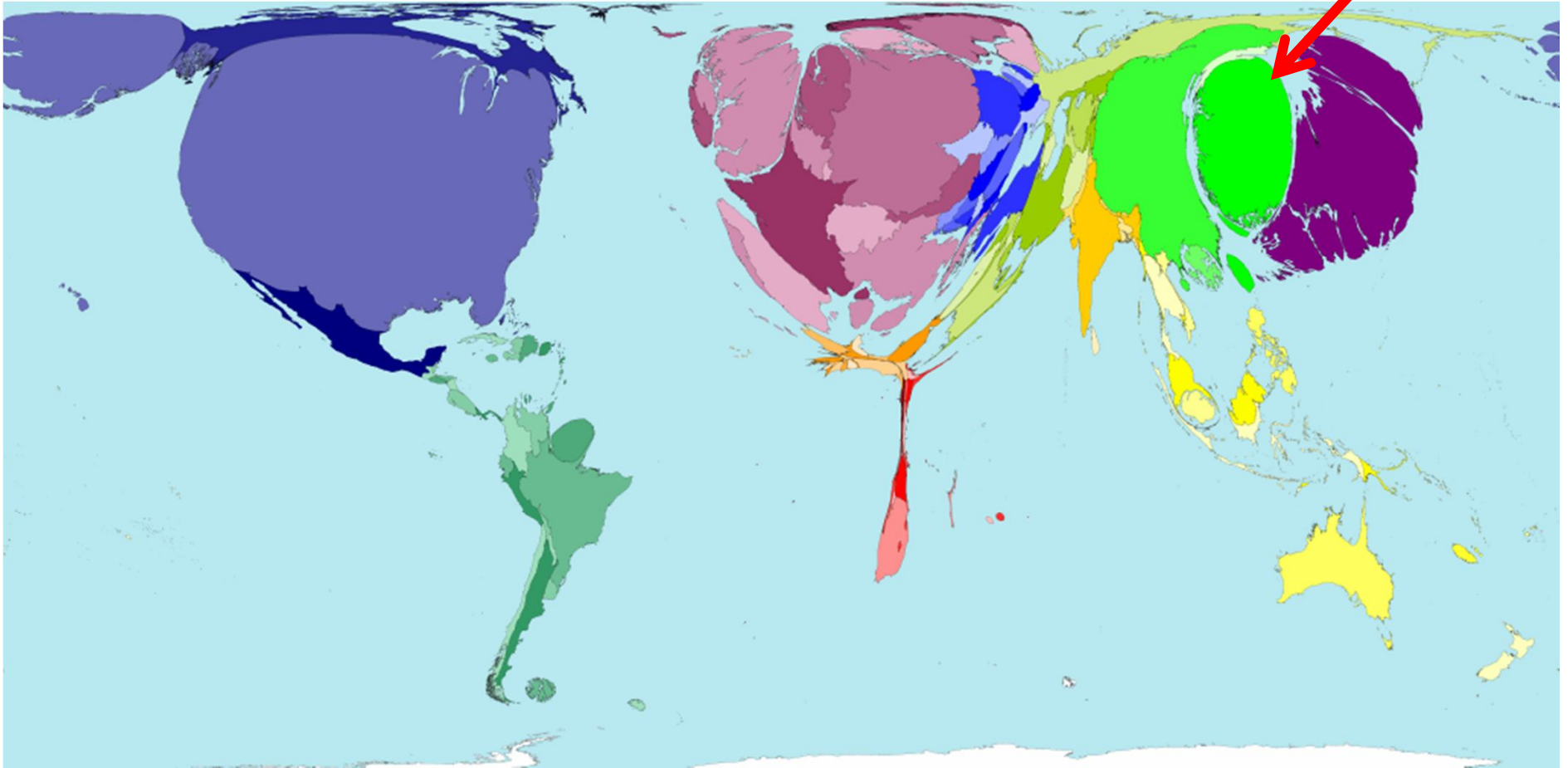
육지면적

Quiz: What map is this?

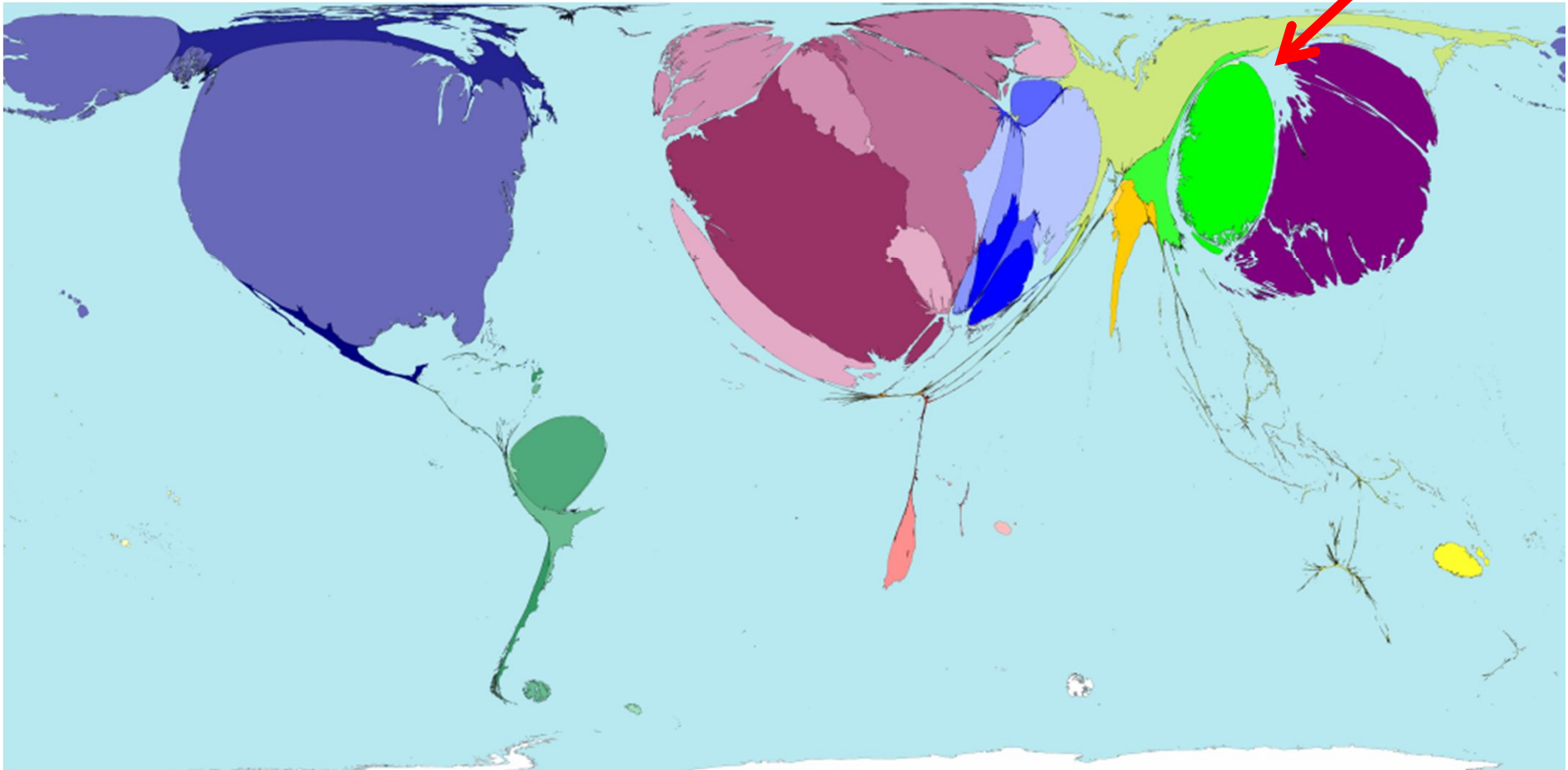


재앙으로 죽는사람

Quiz: What map is this?

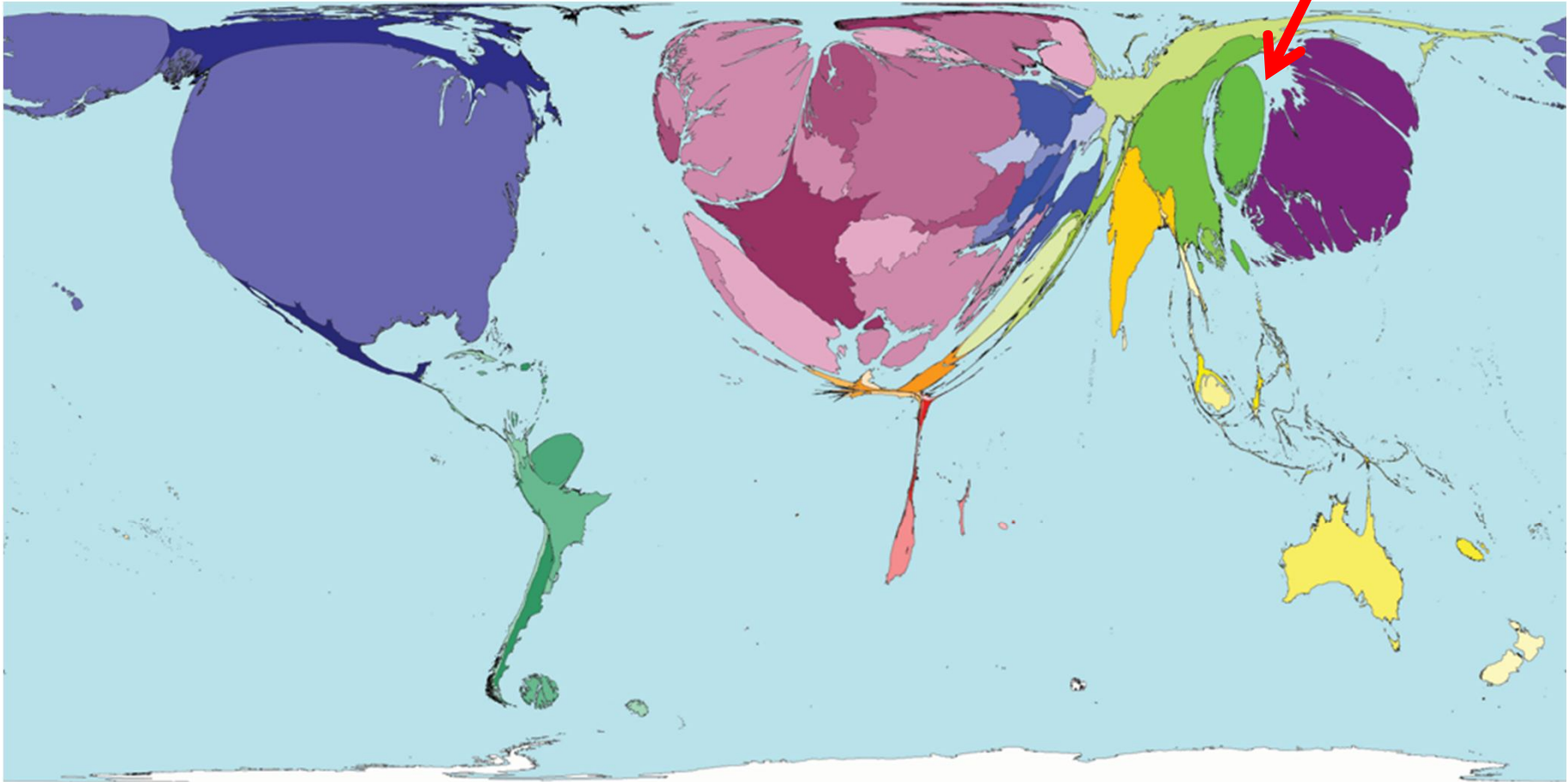


What map?



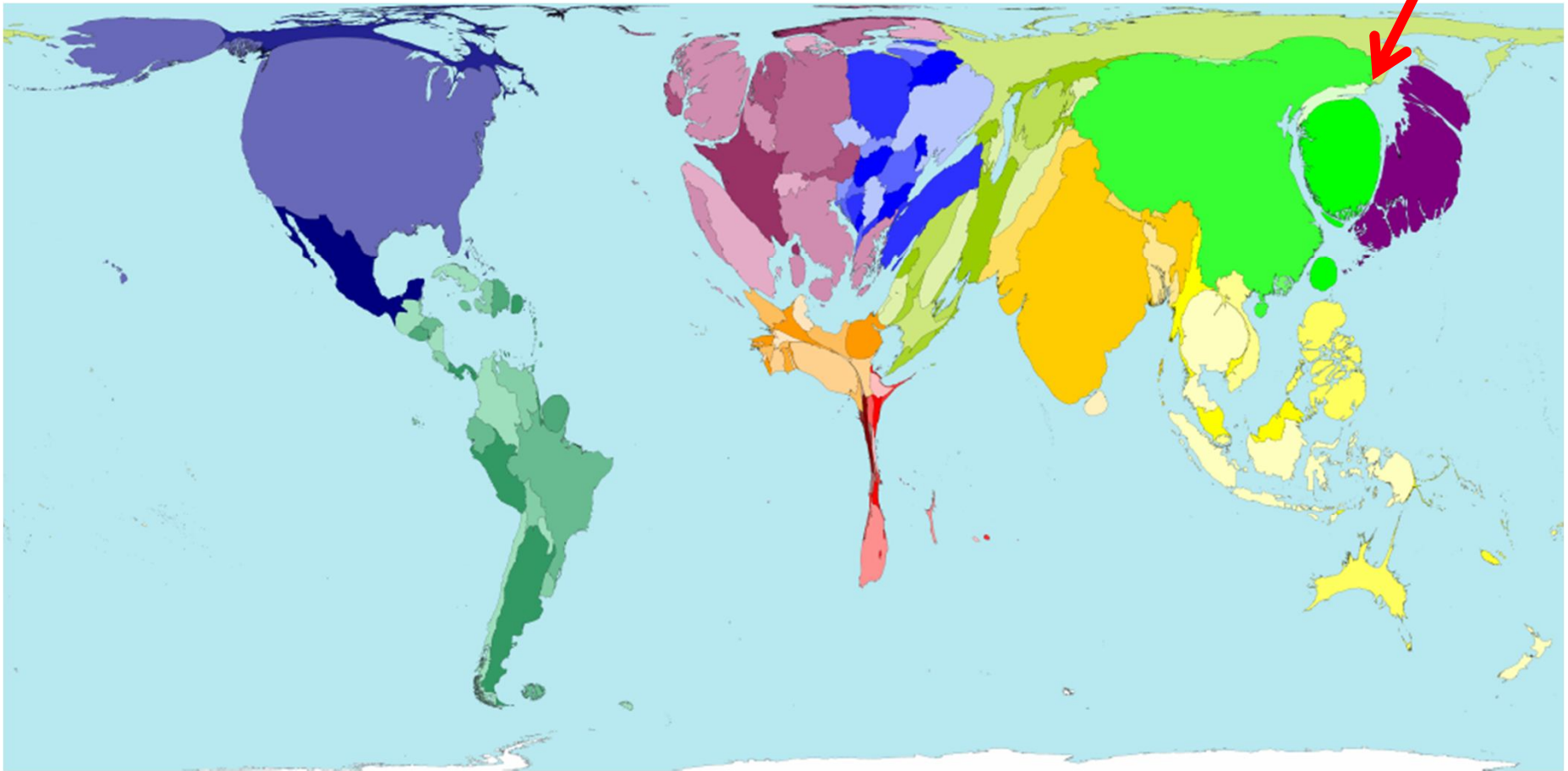
핵연료 사용

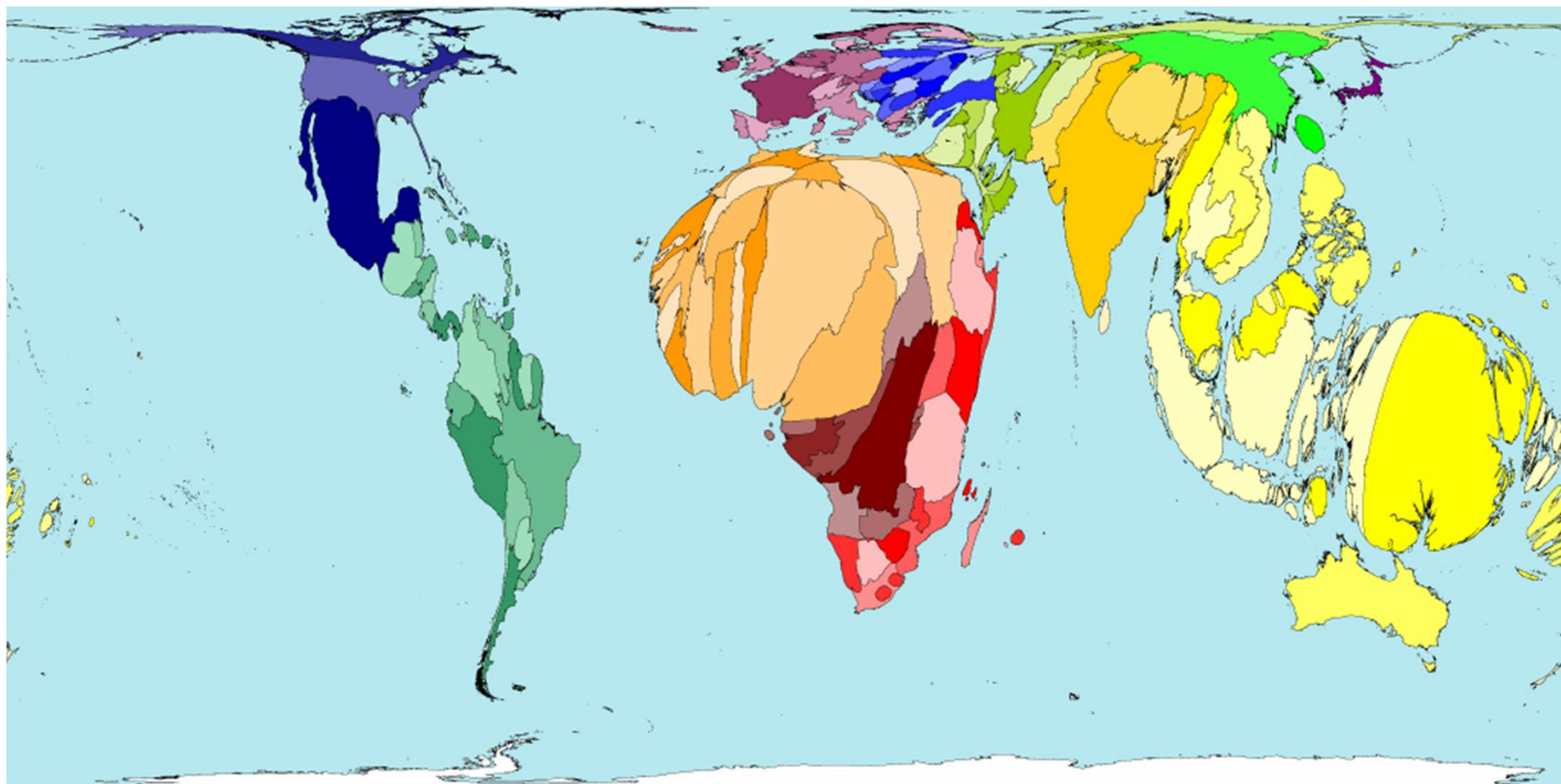
What map?



과학연구

What map?





고유 원주민 언어의 수

Contents

1. Introduction

2. Korean Personal Genome Project

3. OpenKPGP

4. Genome Engine (GiSys)

Conclusion

Let's do more sequencing

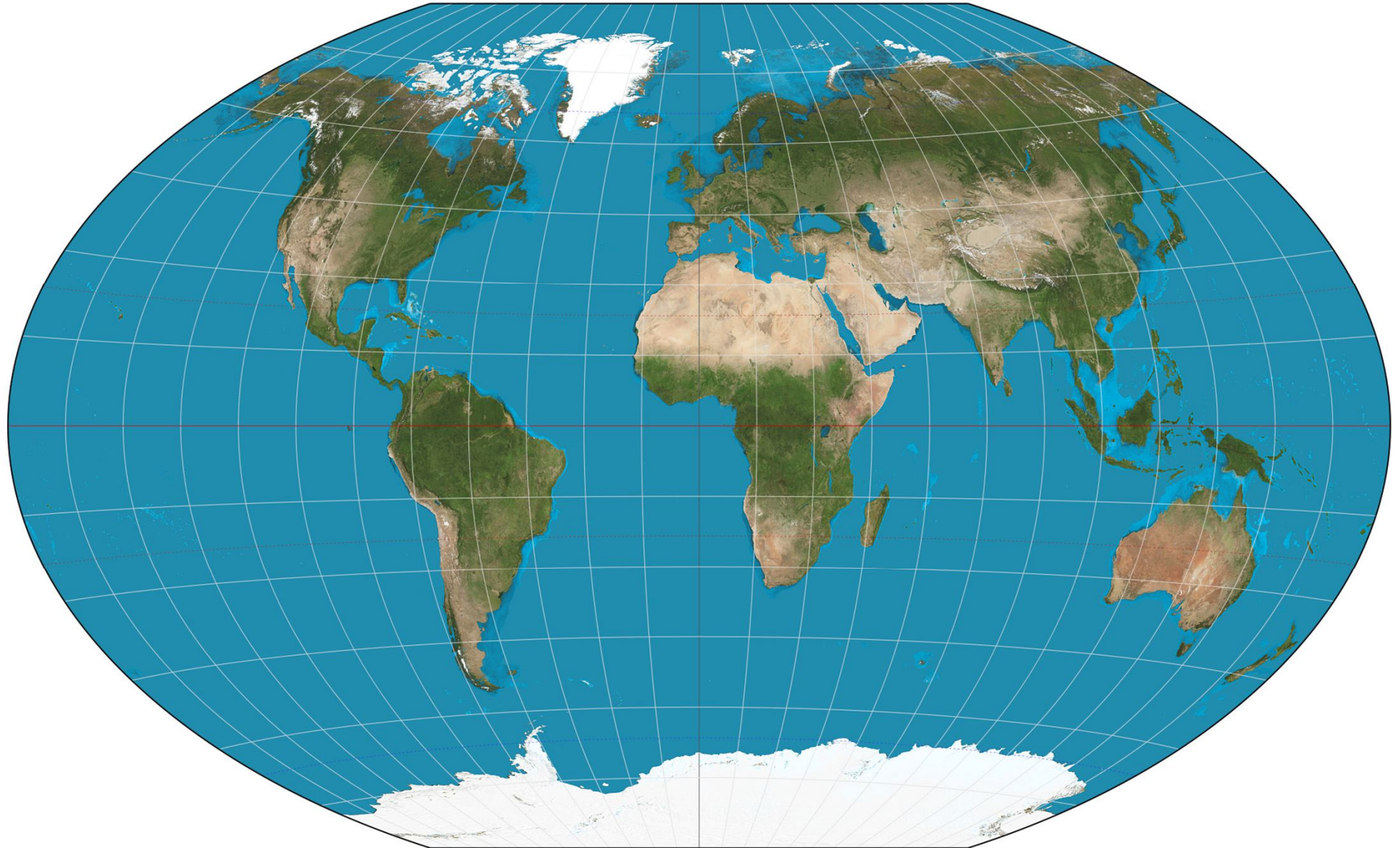
Conclusion

Let's sequence 7 billion human genome

7 billion genome project (7BGP)

The End

BIM: Biological Information Mapping



Bioinformatics Is about Mapping

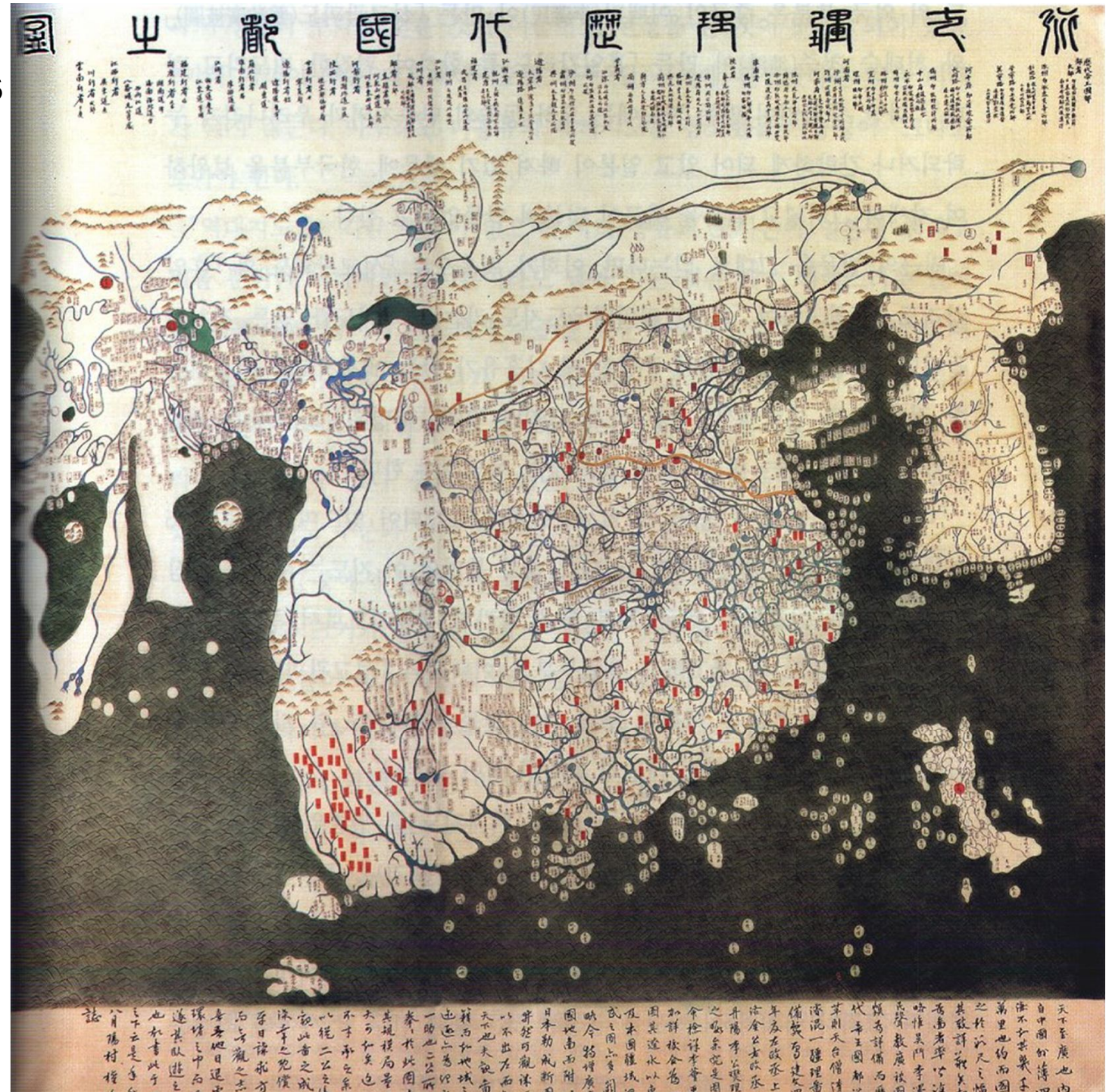
X is IEP
Y is Size

Old map
Is not
Accurate.

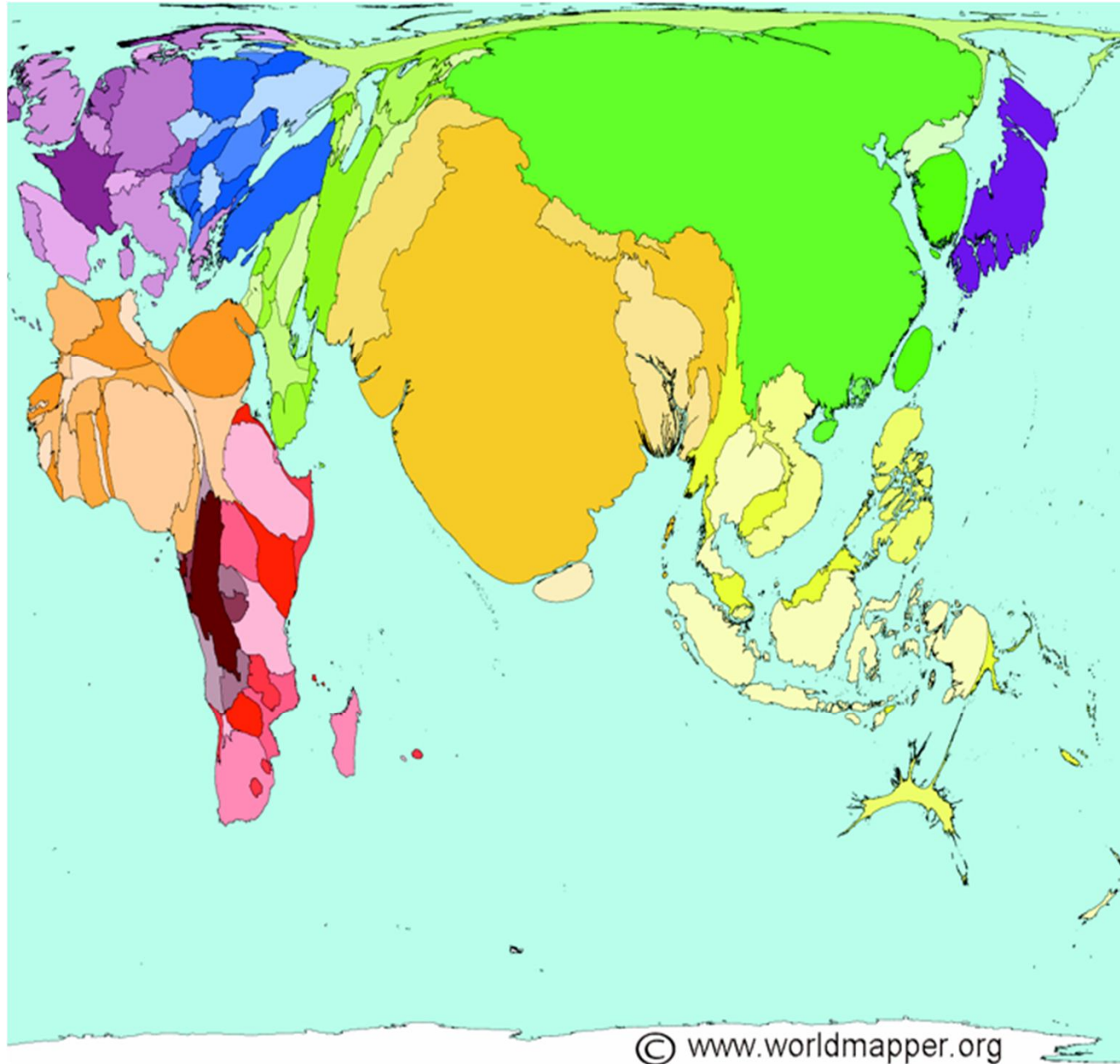
However
It helps
People to
Explore.

Data
Information
Knowledge

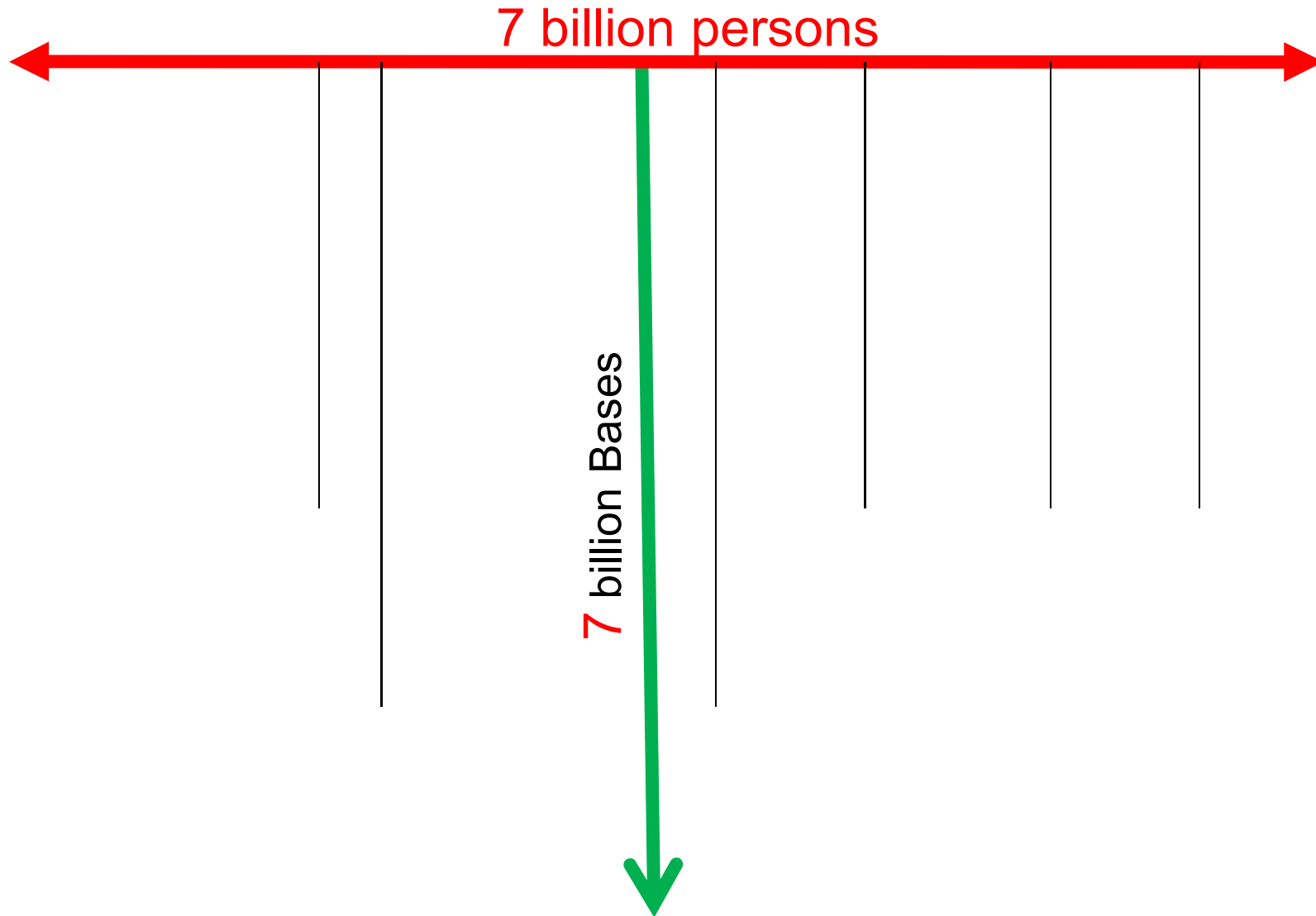
Gangnido 1402



World Population map (2002)

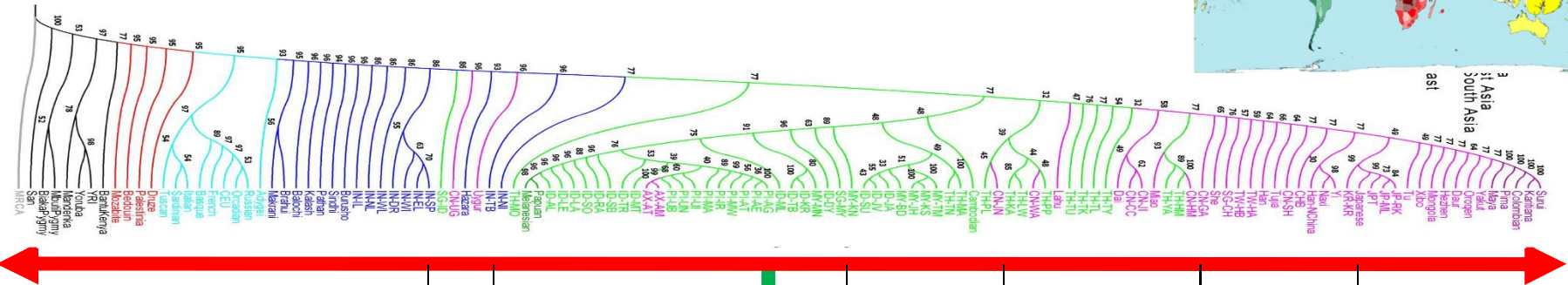
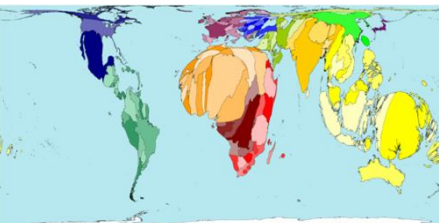


Genomic T map



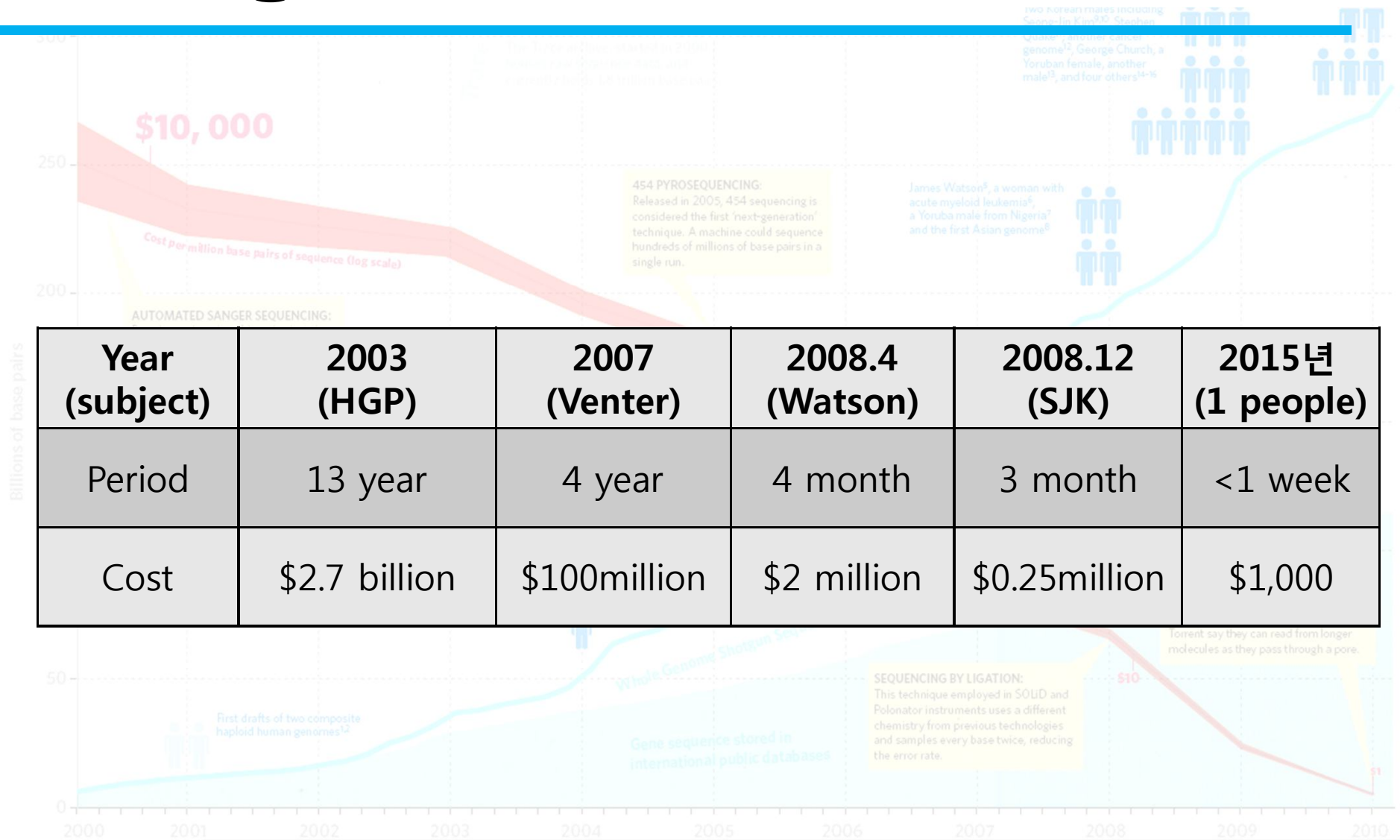
Genome Diversity and Genome Variation

Map: Perception



50,000 Bases for PASNP

Background – Cost & Period



Personal Genome analysis

○ 2008.12 SJK personal genome



Misha Angrist



George Church



Esther Dyson

○ 2008.11 Chinese Personal genome



Jay Flatley



Henry Louis Gates



Rosalynn Gill

○ 2008.4 J. Watson Personal genome



Seong-Jin Kim



Greg Luoler



James Lupski

○ 2007 C. Venter Personal genome

○ 2006 PGP(Personal Genome Project)



Stephen Quake



Dan Stoloescu



James Watson

○ 2003 Human Reference Genome

Personal Genome Data

2011 year

Whole Genome Sequencing
Analysis and Publically Available :
about 300 people

The first Korean Genome Sequence

Downloaded from genome.cshlp.org on August 11, 2009 - Published by Cold Spring Harbor Laboratory Press



The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group

Sung-Min Ahn, Tae-Hyung Kim, Sunghoon Lee, et al.

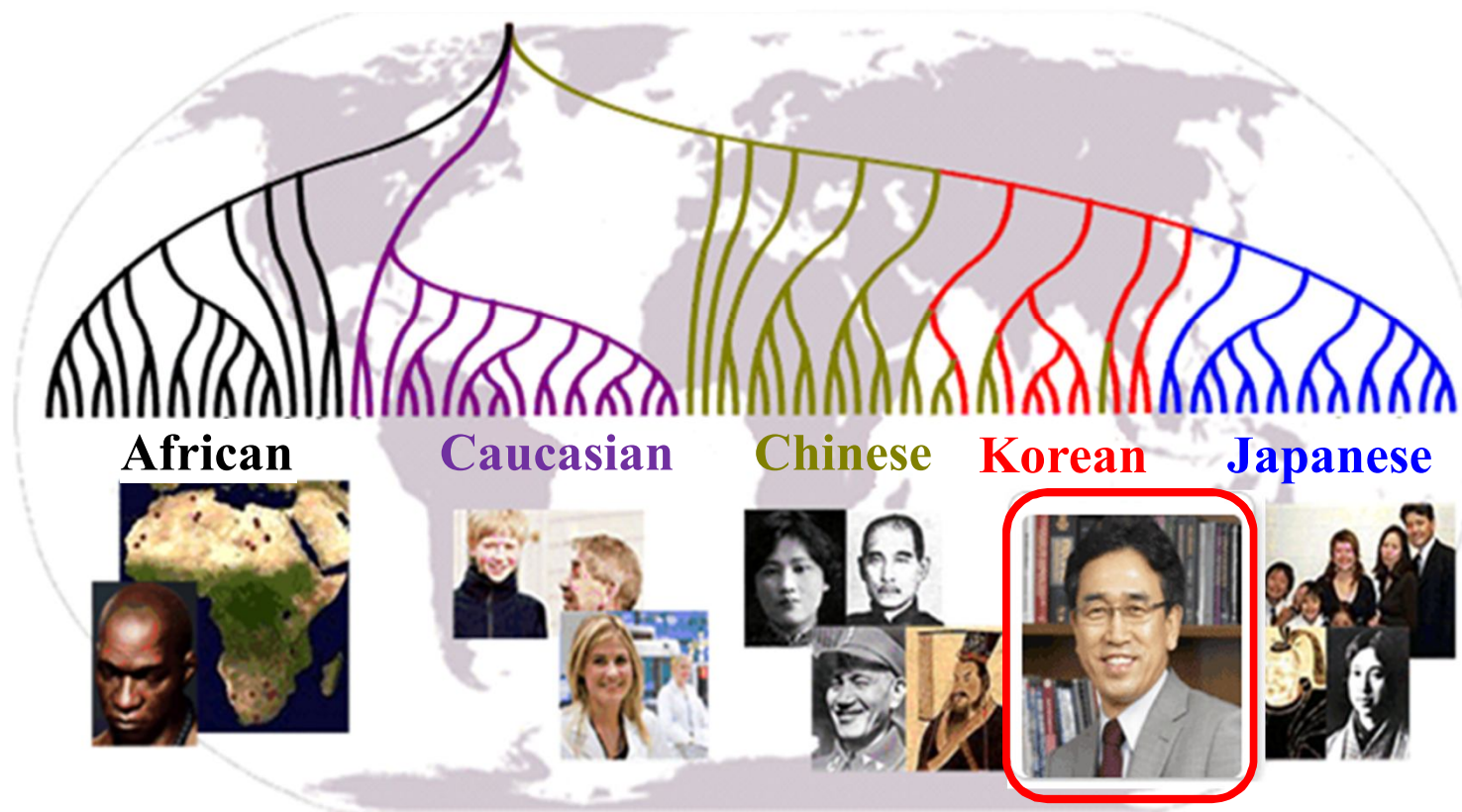
Genome Res. published online May 26, 2009

Access the most recent version at doi:[10.1101/gr.092197.109](https://doi.org/10.1101/gr.092197.109)

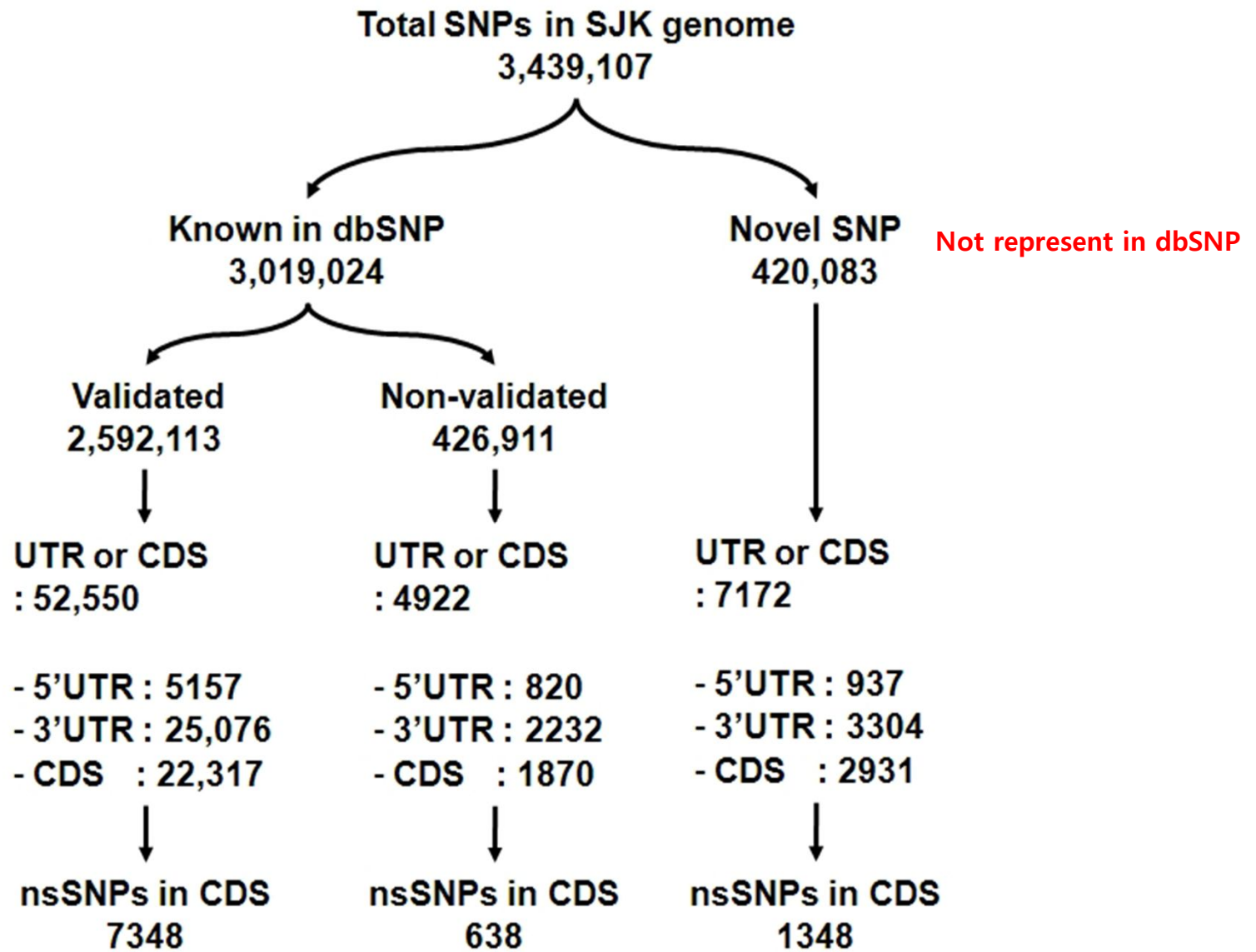
Data publicized: 2008. Dec. <ftp://bioftp.org>

The first Korean Genome analysis

Common Ancestor



Classification and number of intra-genic SNPs





ABOUT US
KPGP 소개

RESEARCH
연구분야

JOIN
참여안내

GENOME'S TALK
유전체이야기

HOME MYPAGE

Community

KOREAN PERSONAL GENOME PROJECT

KOREAN Personal Genome Project

GENOME PROJECT

주요 KPGP
참여현황

KPGP의 성장 과정을
보여드립니다.



JOIN
참여하기

» 누구나 KPGP와 함께 하실 수 있습니다.



유전정보참여



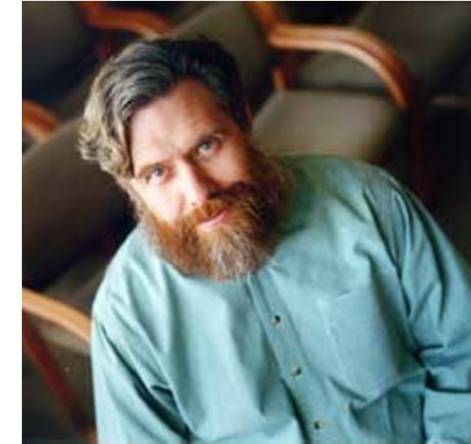
연구 참여



기업참여

Personal Genome Data

- Public Open Source Genome Project
- Volunteers from the general public working together with researchers to advance personal genomics
- Led by Prof. George Church in Harvard Medical School
- **100,000** informed participants from the general public (US Citizen)
- Research Data freely available to the public



Personal Genome Project

Personal Genome Project

Home Project Overview Participation Overview PGP Community [DONATE](#)

Volunteers from the general public working together with researchers to advance personal genomics.

We believe individuals from the general public have a vital role to play in making personal genomes useful. We are recruiting volunteers who are willing to share their genome sequence and many types of personal information with the research community and the general public, so that together we will be better able to advance our understanding of genetic and environmental contributions to human traits. Learn more about how to [participate](#) in the Personal Genome Project.

Project Overview. The PGP hopes to make personal genome sequencing more affordable, accessible, and useful for humankind. Learn more about our [mission](#).

Want to participate? We aim to enroll 100,000 informed participants from the general public. Learn more about [participation](#) in the PGP and how you can get involved.

Meet our volunteers. Participants may volunteer to publicly share their DNA sequence and other personal information for research and education. Meet the "PGP-1K".

Documentary Film about PGP. Two-time Emmy Award-winning documentary producer [Marilyn Ness](#) is making a film about the PGP. [Watch webisode 1, 2, and 3.](#)

CC0. We are committed to making [research data](#) from the PGP freely available to the public. Read about PersonalGenomes.org's use of the [CC0 universal waiver](#).

GET Conference. Hear from thought leaders working at the frontiers of personal genomics in Philadelphia at the 2011 [Genomes, Environments, and Traits Conference](#).

Participant Login
[Login Now](#)

Project News
[Subscribe to our newsletter.](#)

Oct 5, 2011: PGP-HMS prepares for national blood collection campaign, adds hundreds of walk-in clinics to network. [See list.](#)

Sep 10, 2011: KPGP publishes 32 genomes of Korean participants. [More.](#)

July 15, 2011: PG.org awarded grant from Google to develop privacy-enhancing tools to enable individuals to donate medical records to scientific research.

April 27, 2011: The GET Conference 2011 was held in Philadelphia at UPenn's new Translational Research Center. [Learn more.](#)

[Sign-up](#), [login](#), [donate](#), or read more [news](#).

Sep 10, 2011 **KPGP publishes 32** genomes of Korean participants. [More.](#)



KPGP Mission

To build a **standard Korean genetic information** database

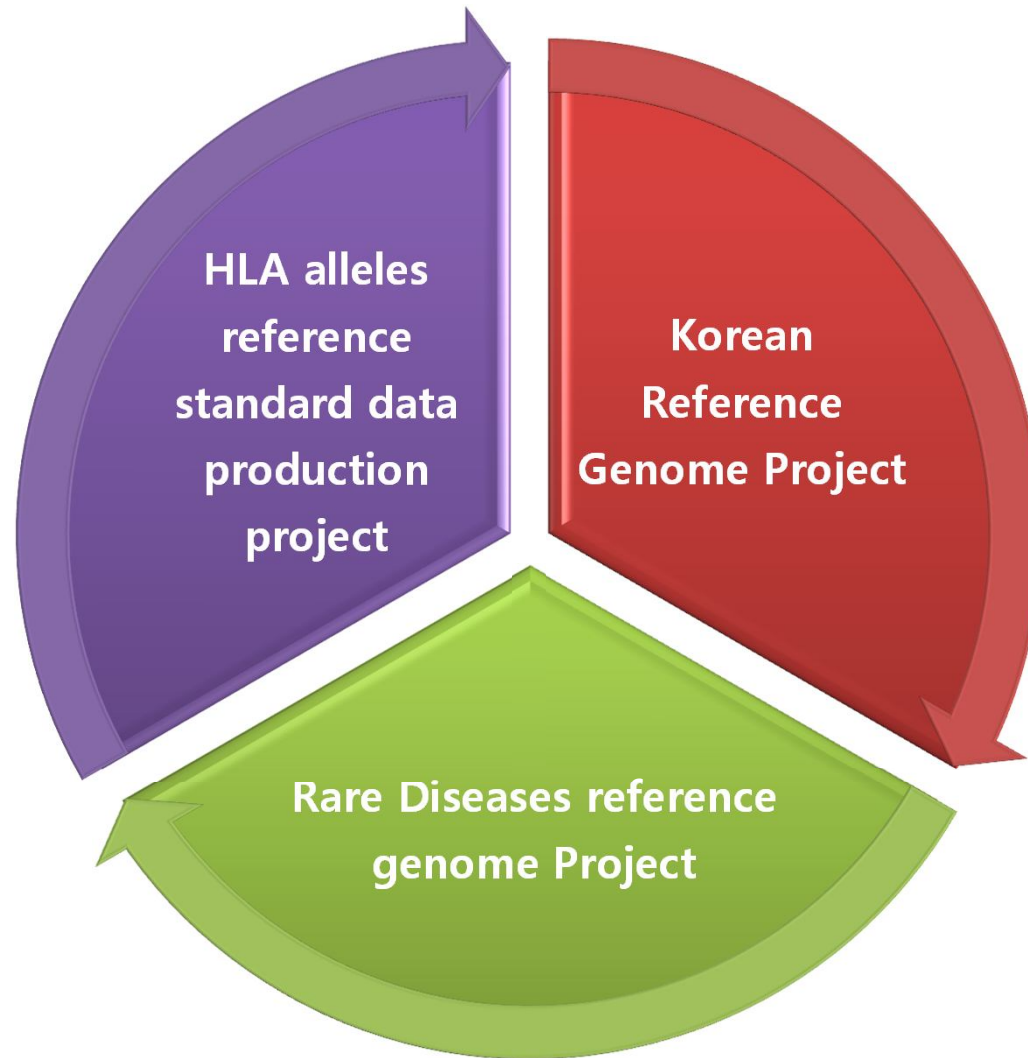
To promote **personalized and personal genome** research and commercialization



To develop **genome sequencing and analysis technologies** in the age of personal genetic information services

To help **community be aware of ethical, social, and legal issues** in regard to genetic information

Research field



KPGP-20

Thanks to KT (Korea Telecom)

Population Genomics Project



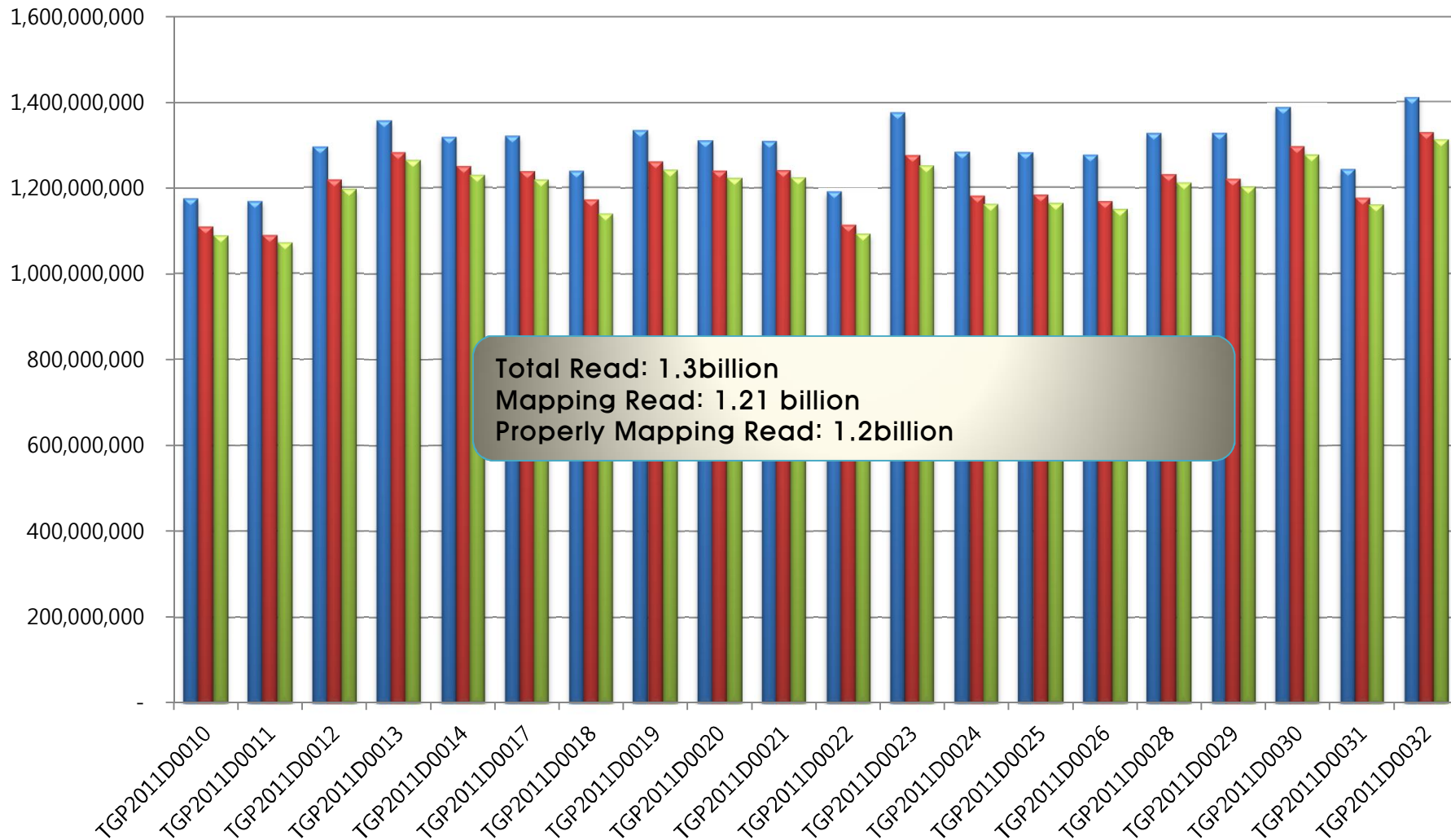
KPGP-20 Result 1

	1 person (First Korean genome)	20 people	Percentage change
Announcement	12 /2008	9 /2011	-
Genome Analysis Duration	3 month	1 month	3-fold decrease ↓
Genome Analysis Cost (per person)	\$0.25million	\$0.6million (\$30,000)	8-fold decrease ↓
Detected Variation (Removed duplication)	3.43M	71.32M (8.76M)	20-fold increase ↑
Novel Variation (Removed duplication)	420K	4.24M (1.84M)	10-fold increase ↑
Read Length	36bp, 75bp	90 bp	2-fold increase ↑
Depth	28.73	37.11	0.3-fold increase ↑

KPGP-20 Result 2

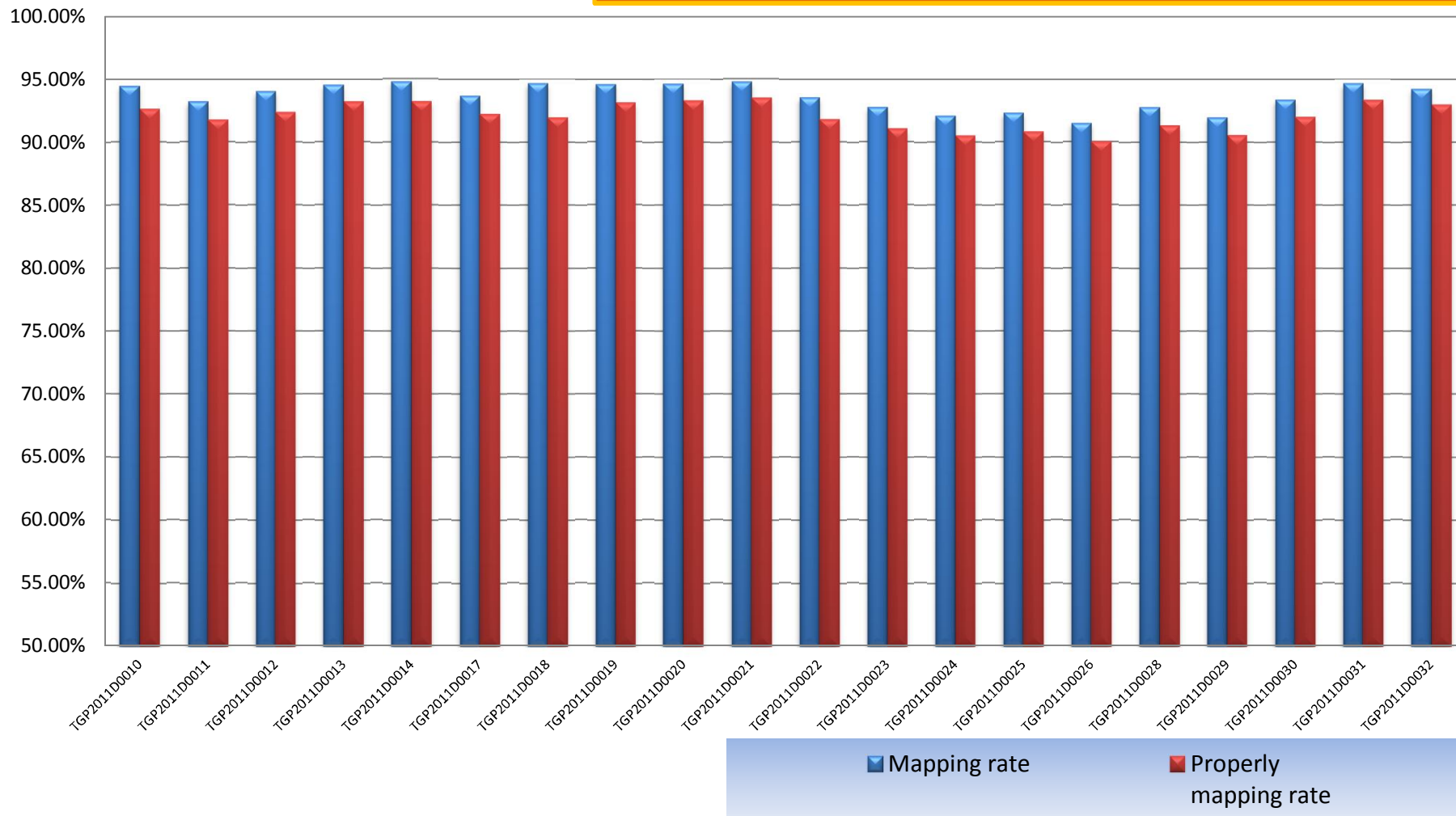
Sample	Raw data Size	Total_reads	Mapped_reads	Mapping_rate	Properly mapped_reads	Properly mapping rate	Depth	Coverage
TGP2011D0010	85Gb	1,174,786,230	1,109,519,944	94.44%	1,088,051,304	92.62%	33.97	98.67%
TGP2011D0011	86Gb	1,167,798,192	1,088,172,839	93.18%	1,071,377,522	91.74%	33.23	98.67%
TGP2011D0012	94Gb	1,295,848,662	1,218,097,132	94.00%	1,197,064,878	92.38%	37.13	99.41%
TGP2011D0013	101Gb	1,358,192,924	1,283,956,209	94.53%	1,265,733,484	93.19%	39.66	98.67%
TGP2011D0014	98Gb	1,318,559,622	1,250,065,225	94.81%	1,229,266,548	93.23%	38.43	99.33%
TGP2011D0017	97Gb	1,320,926,164	1,237,634,224	93.69%	1,218,621,568	92.26%	37.75	99.23%
TGP2011D0018	90Gb	1,238,984,198	1,173,023,066	94.68%	1,139,456,378	91.97%	36.25	98.66%
TGP2011D0019	99Gb	1,334,920,326	1,262,381,805	94.57%	1,242,631,260	93.09%	38.56	99.36%
TGP2011D0020	97Gb	1,311,413,946	1,241,084,671	94.64%	1,223,679,352	93.31%	38.26	98.69%
TGP2011D0021	96Gb	1,309,436,430	1,241,318,554	94.80%	1,224,736,606	93.53%	38.42	98.67%
TGP2011D0022	86Gb	1,190,268,342	1,113,253,155	93.53%	1,092,477,330	91.78%	34.00	99.39%
TGP2011D0023	101Gb	1,376,317,182	1,276,555,687	92.75%	1,252,845,038	91.03%	38.97	98.68%
TGP2011D0024	94Gb	1,282,791,040	1,181,054,915	92.07%	1,161,234,780	90.52%	35.58	99.33%
TGP2011D0025	93Gb	1,281,863,526	1,182,955,980	92.28%	1,163,771,886	90.79%	35.75	99.39%
TGP2011D0026	94Gb	1,277,607,666	1,168,806,522	91.48%	1,149,977,950	90.01%	35.28	99.44%
TGP2011D0028	97Gb	1,327,623,402	1,232,059,676	92.80%	1,212,545,772	91.33%	37.69	98.82%
TGP2011D0029	97Gb	1,328,229,348	1,221,368,842	91.95%	1,202,909,240	90.56%	37.35	98.83%
TGP2011D0030	103Gb	1,389,159,180	1,297,117,533	93.37%	1,277,997,460	92.00%	39.42	99.44%
TGP2011D0031	89Gb	1,242,908,982	1,176,511,586	94.66%	1,160,494,548	93.37%	35.89	99.42%
TGP2011D0032	105Gb	1,411,123,640	1,329,762,106	94.23%	1,312,030,388	92.98%	40.51	99.39%

KPGP-20 Result 3



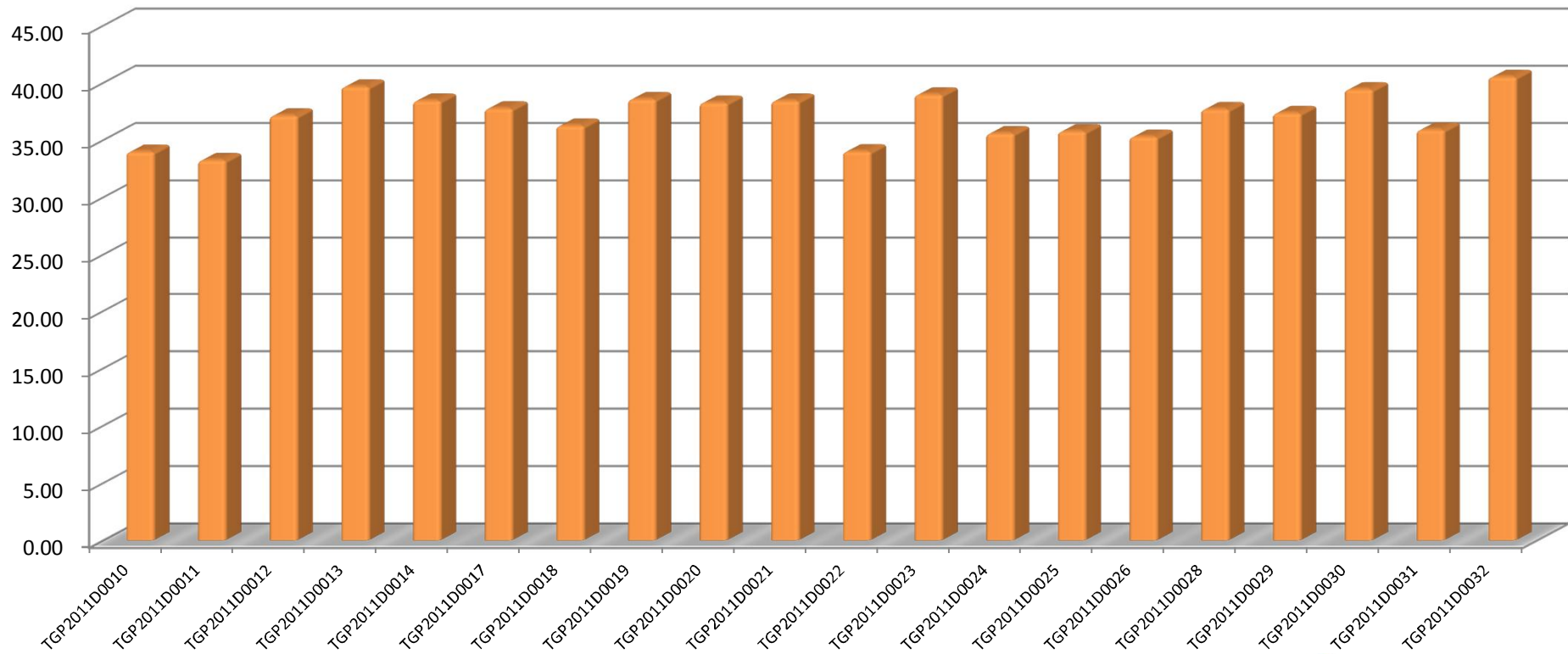
KPGP-20 Result 4

Average Mapping Rate: 93.62%
Average Properly Mapping Rate: 92.08%



KPGP-20 Result 5

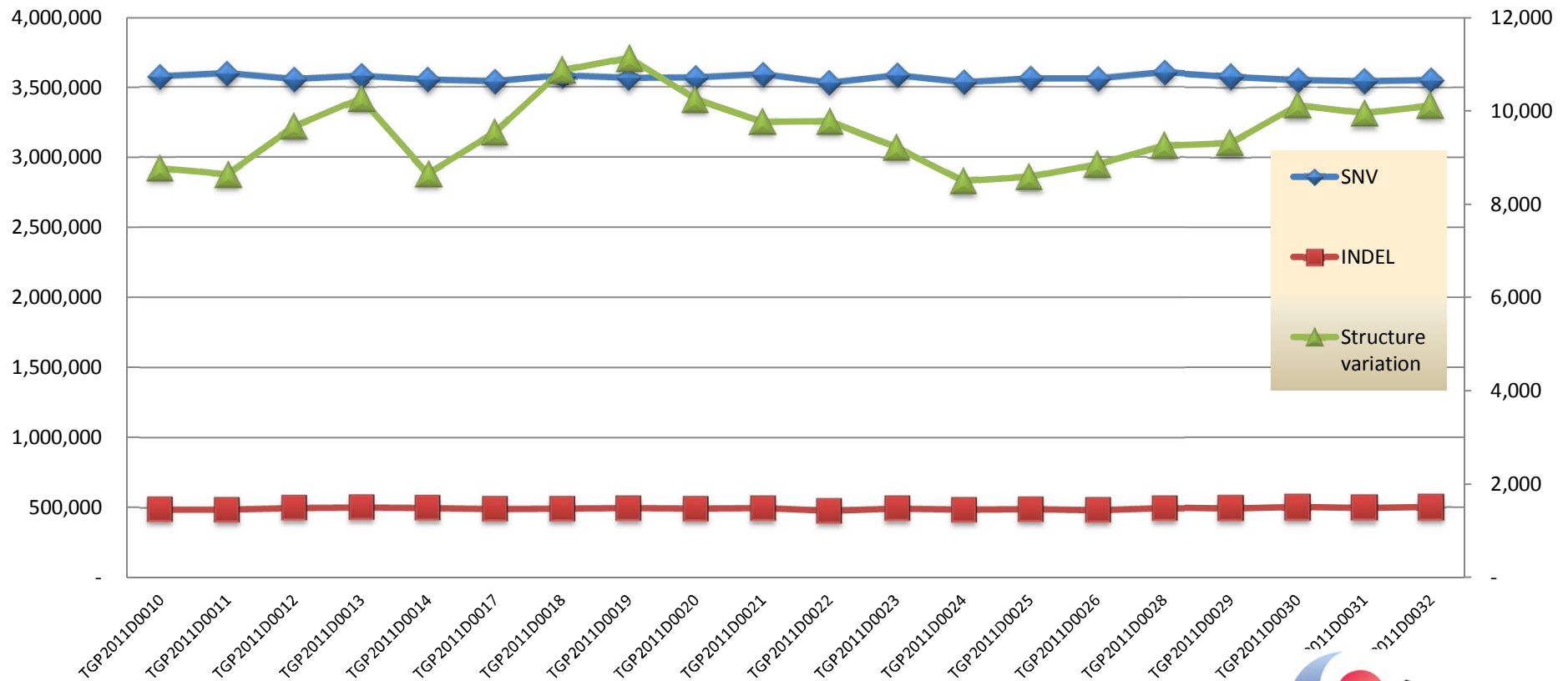
Average Depth: 37.11X



KPGP-20 Result 6

Structural variation

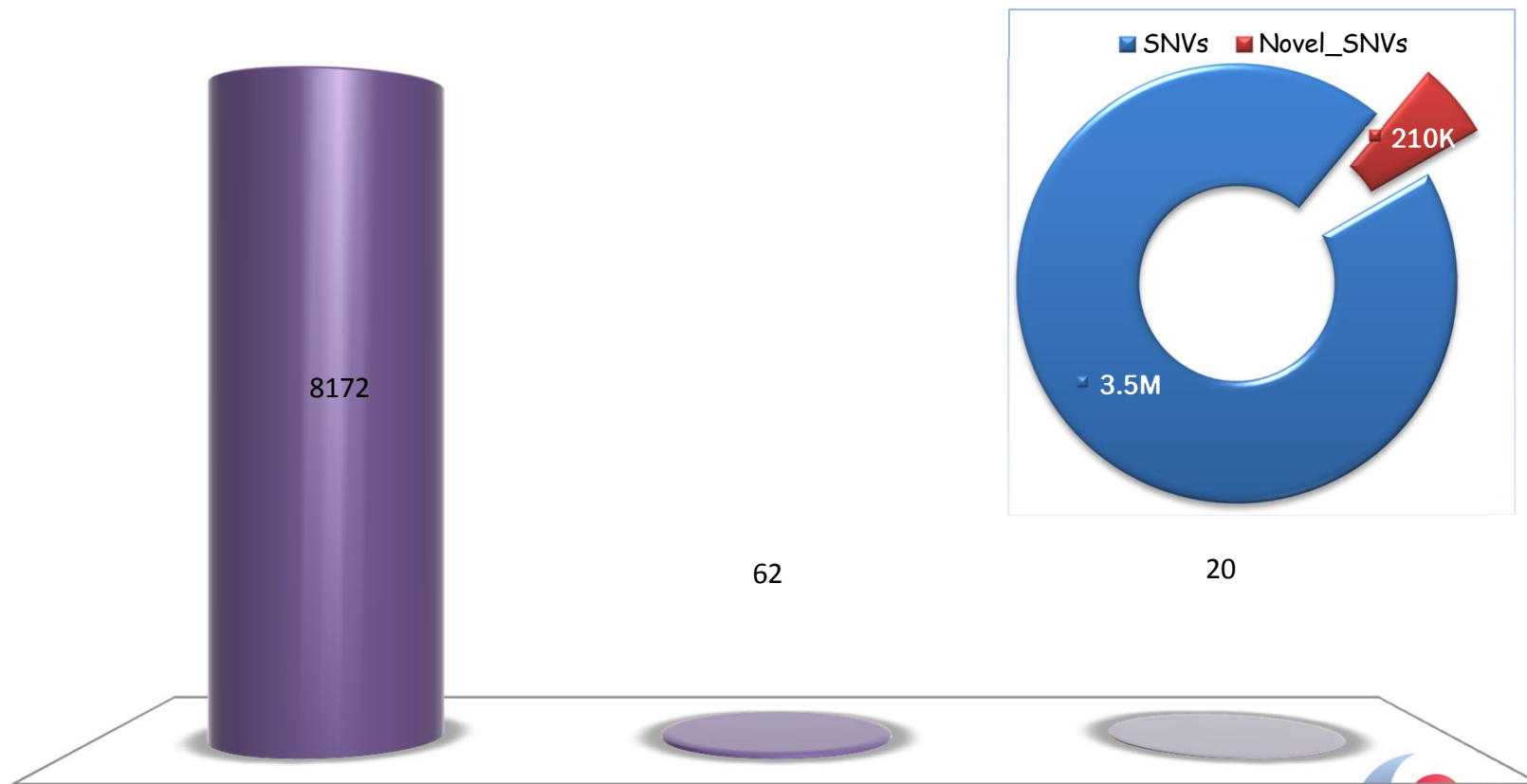
Average SNV: 3.5M
Average Indel: 490K
Average SV: 9,500



KPGP-20 Result 7

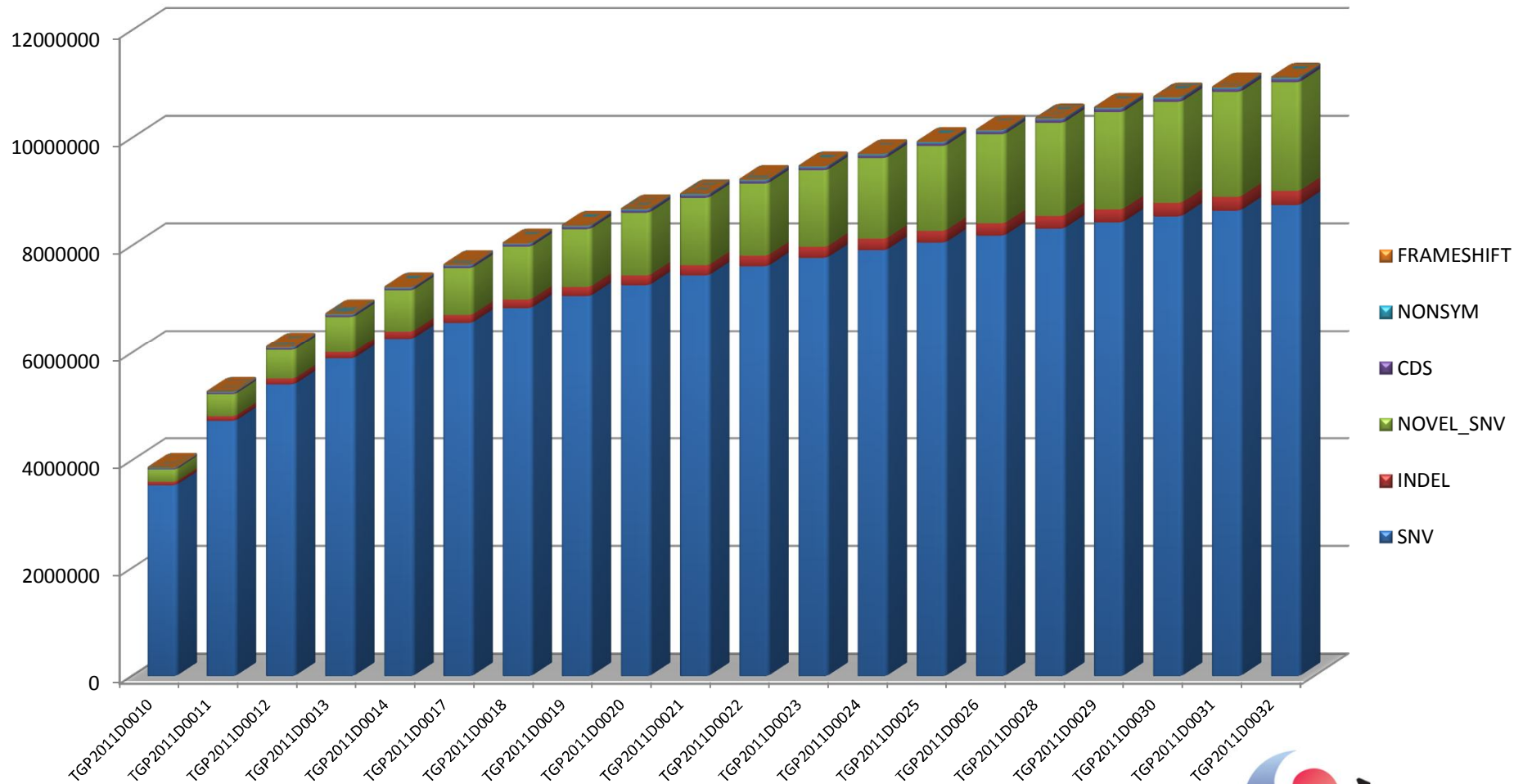
Individual Average SNVs

- Non Synonymous missense
- Non Synonymous Nonsense
- Frameshift

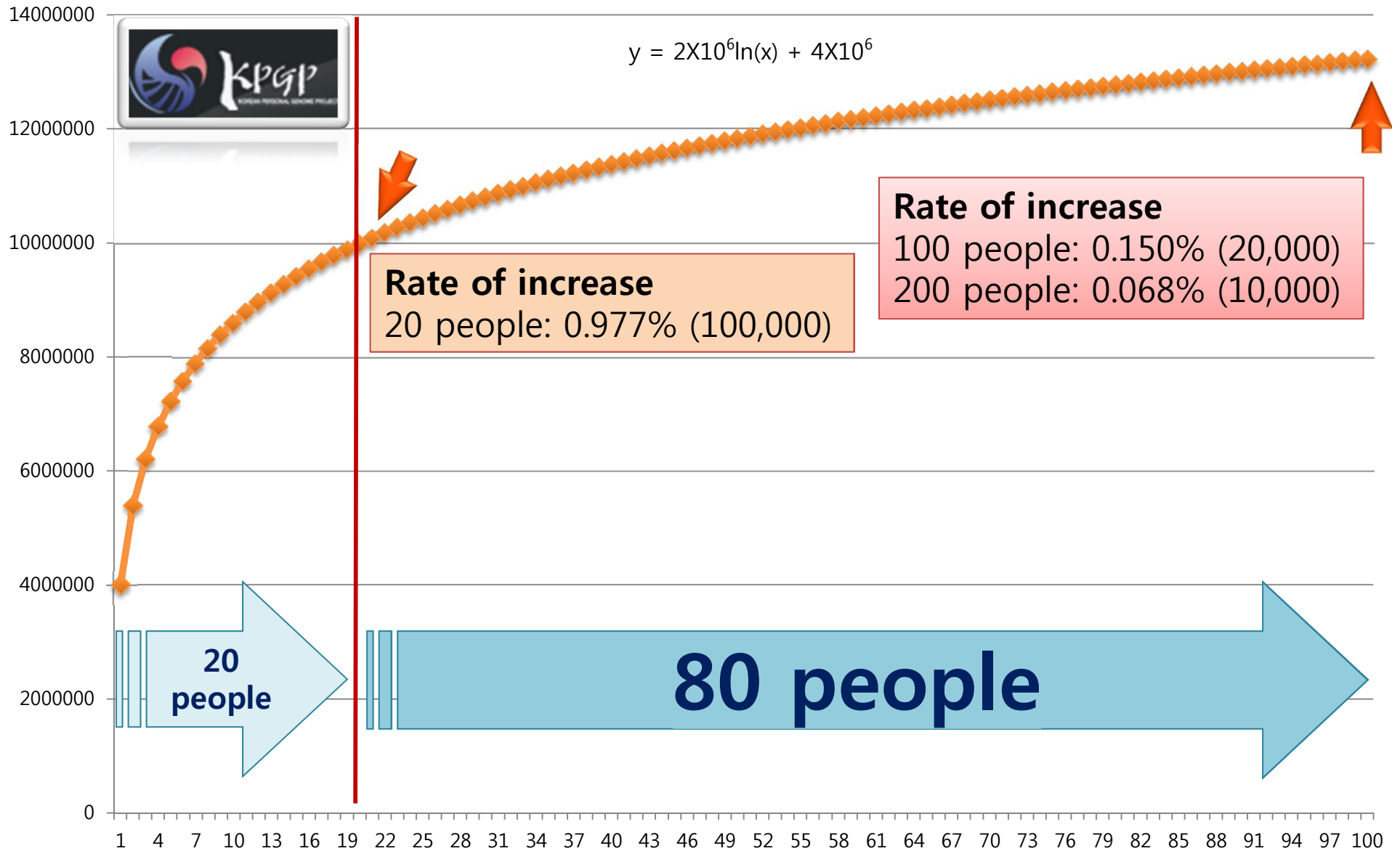


KPGP-20 Result 8 / 10

Cumulative of Genetic Variant



KPGP-20 Result 9 / 10



KPGP-20 Result 10

Representative genetic variants of Korean

dbSNP 132 Variants

KPGP-20 Variants

26.1M
variants

6.91M
Variants

1.84M
Variants

60K
Variants

**Discovery of newer variants in
50% of KPGP-20 samples**

OPEN KPGP



OPEN KPGP

OPEN KPGP

Korean Participation of **PGP**

'**OPEN KPGP**' Launched by Genome Research Foundation(GRF)

Providing Genome Data to Researchers **without Restrictions**

Korean Personal Genome Project (KPGP) is a participative research project established by Genome Research Foundation(GRF).

The international Personal Genome Project (PGP), led by U.S. none-profit research group, is founded in 2006 and seeking a diverse range of volunteers with the purpose of improving human health.

In concordance with the international PGP, KPGP has started to improve Korean people's health and medical welfare. > [more about us](#)

KPGP Public Data



Korean



Other



Rare Disease Patients



Multiracial(people)



Monozygotic Twin

olleh ucloud

KTuCloud+PGI



Dizygotic Twin



OPEN KPGP - Data

KPGP 00001

- Public Profile ID : KPGP_00001
- Gender : F
- Nationality : Korea
- Ethnic Group : East Asian
- Platform : Illumina,HiSeq2000

BioFTP Download

Torrent Download

Genetic data Download

- Raw Data [Raw Data](#)
- Aligned Data [Mapping](#)
- Nonsynonymous SNVs [Nonsynonymous SNVs](#)
- Read me [Read me](#)
- Genome Variation [SNVs](#) [CNVs](#) [INDELs](#) [SVs](#)

Torrent Download

One Data File Format (ODFF) Reader recommended by Genome Research Foundation

File name : KPGP1_G_110915_HiSeq_EastAsian_Kor_F

Sample ID : KPGP_00001

Other IDs : TGP201000010

Sample type : Genomic DNA

Tissue type : Blood

Data Publication Date: 2011.9.15

Sequencing platform : Illumina, HiSeq2000

Ethnic Group : East Asian

Nationality : Korea

Gender : F

Brith Date : N/A

Diseases : N/A

Name : N/A

Data depository: <ftp://bioftp.org/B10/Distribute/Open-KPGP/TGP201000010/>

Project name: KPGP






Reference: UCSC, HG19

Contact email: jongbhak@yahoo.com

Data analysis tool information

- Rawdata fastq
- mapping result bwa(0.5.9)
- SNV santools(0.1.16)
- INDEL santools(0.1.16)
- SV breakdancer (1.1)
- NSSNV GRF's pipeline
- CNV N/A
- STATISTICS GRF's pipeline

OPEN KPGP – Available Data

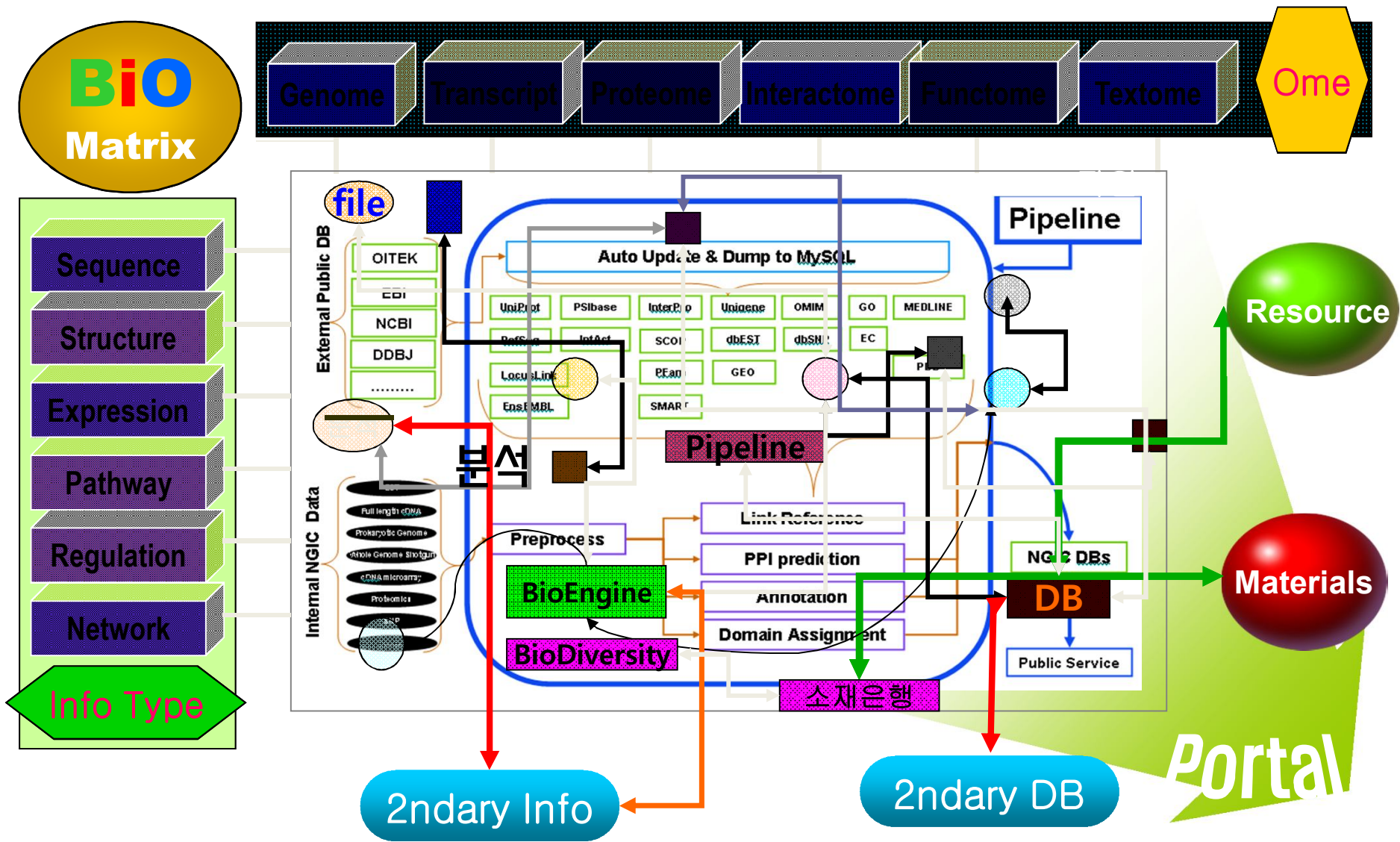
Korean	 Korean	35 people
Multiracial	 Multiracial(people)	2 people
Caucasian	 Other	1 person
Monozygotic twin	 Monozygotic Twin	2 people
Dizygotic twin	 Dizygotic Twin	2 people
Publically available samples		38 people

Genome Engine (GiSys)

Korean National NGS software project

Ministry of Economy and Knowledge

BioEngine: Putting structure in omics



GiSys Service Business



Cloud Management Interface

Administration and Management

Service Request Examiner and Admission Control

- Customer-driven Service Management
- Computational Risk Management
- Autonomic Resource Management



Pricing

Service Request Monitor

VM Monitor

Accounting

Dispatcher

Bio Application

Genome Annotation System

BioWorkbench

Bio Workflow(Pipeline)

De-nove, Re-sequencing, RNA-Seq, ChIP-Seq
Bisulfite-Seq, MeDIP-Seq ...

Bio Module

BWA

Botie

Samtools

GATK

Picard

Bedtools

FastX toolkit

EMBOSS

Biopieces

...

Bio Middleware

Hadoop - MapReduce

Apache Web Service

Backup

Identity

Object Storage

Distributed File System

Security

Integration

BioOS(Linux)



BioInfra



Servers



Storage

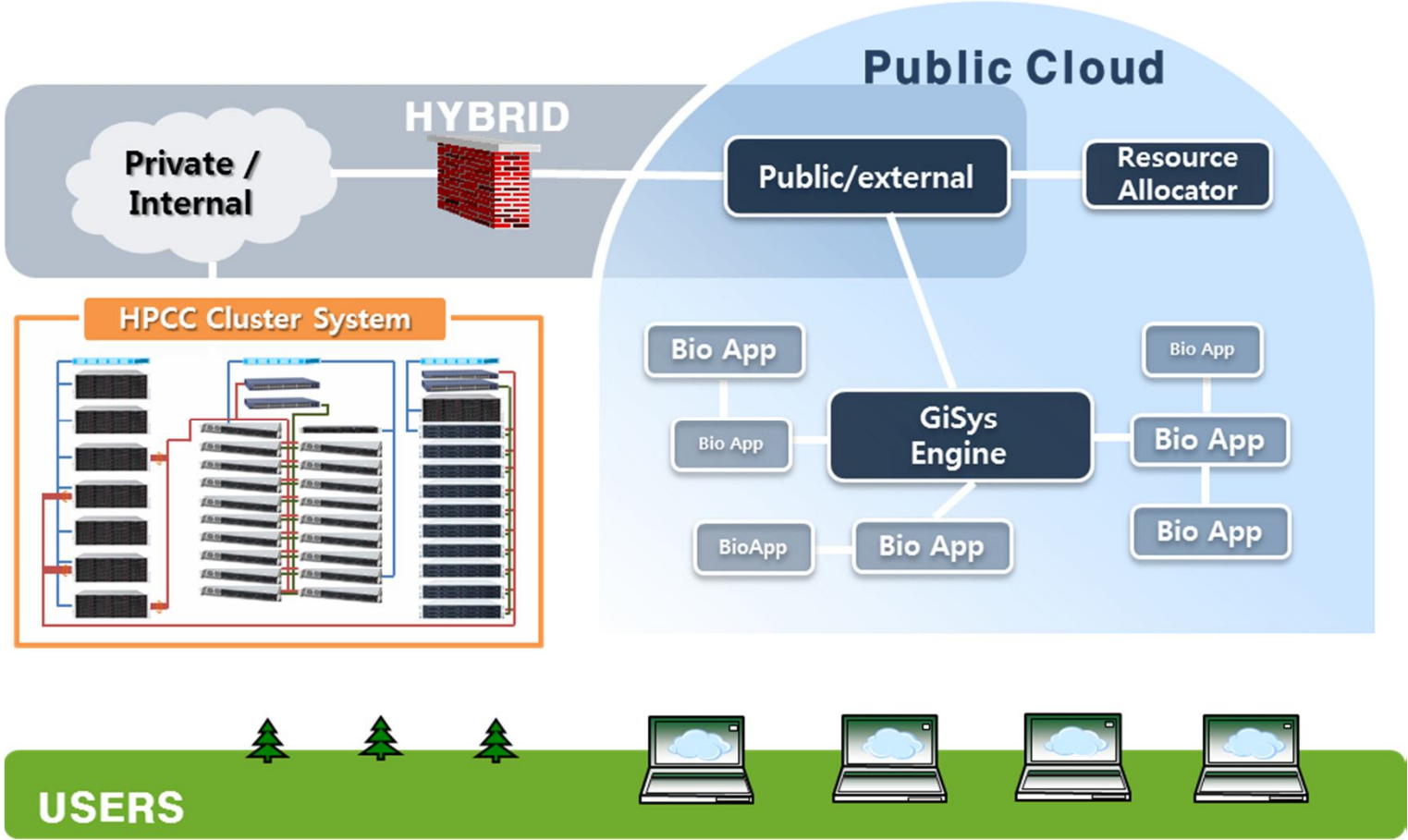


Internal

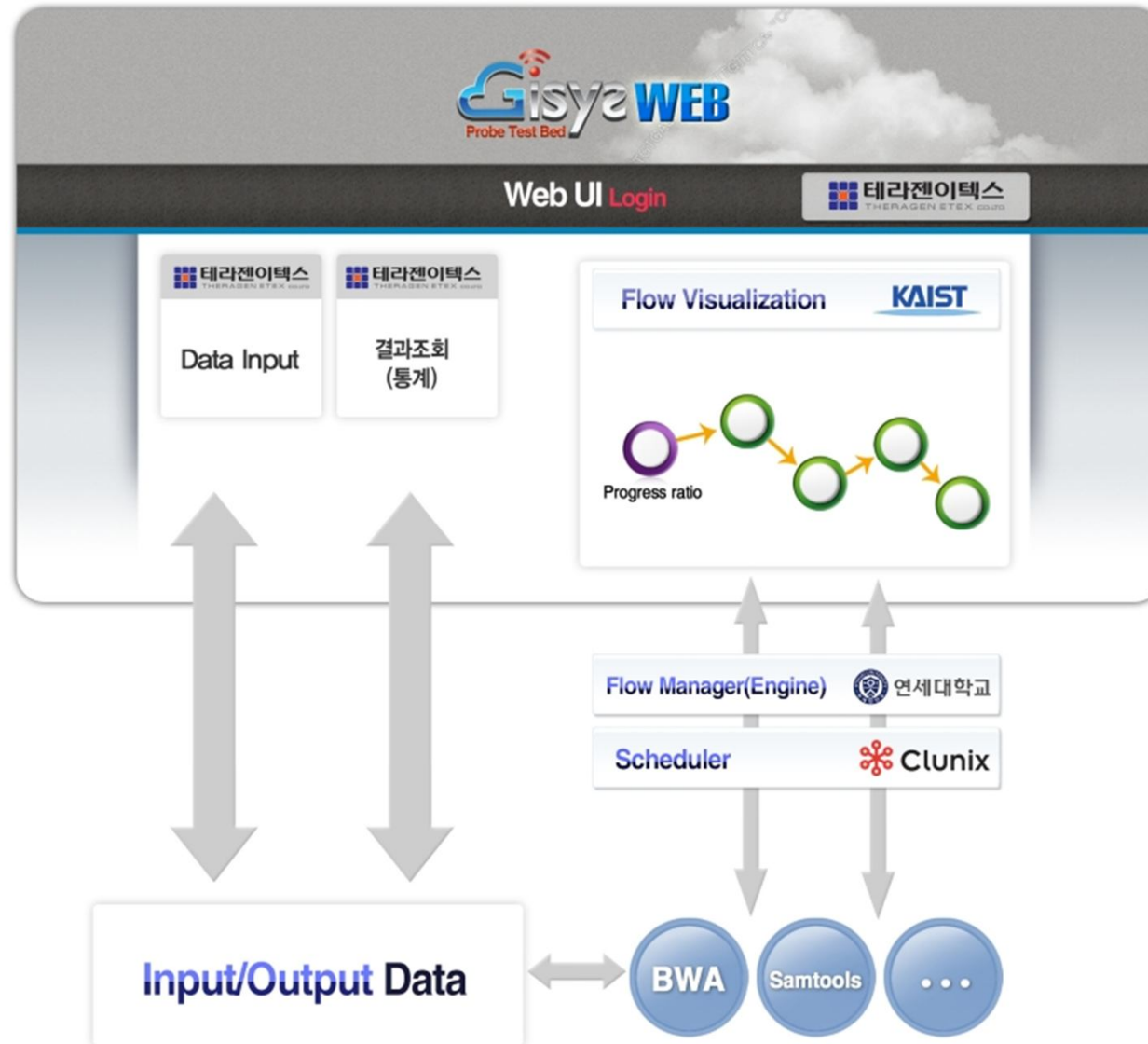


External

GiSys (Genome Informatics System)



GiSys (Genome Informatics System)



GiSys

The screenshot displays the GiSys web interface. At the top, there is a navigation bar with three tabs: 'About', 'Analysis' (which is active), and 'Progress'. Below the navigation bar, the breadcrumb path is 'Home > Analysis > Whole genome analysis'. A notice in Korean states: '※ 위 분석 페이지는 GiSys Proof Test Bed를 위한 페이지이며, 추후 상용화시에 기능 추가 및 변경이 이루어 질 수 있습니다.' Below this, the page title is '>Human Whole Genome Analysis Procedures'. A progress indicator shows four steps: 01 (selected), 02, 03, and 04. A modal window titled 'Human Whole Genome Analysis Pipeline' is open, showing the pipeline steps for GiSys-A1. The steps are: BwaAlign (command: bwa index;bwa aln), BwaPairing (command: bwa sampe), Reformatting (commands: samtools view;samtools sort;samtools index), VariantCalling (command: samtools mpileup), and ReportBuilding (command: make a report). On the left side of the interface, there is a sidebar with links for 'Whole genome analysis', 'CL community', and 'Biomap'. Below the sidebar, there are partially visible labels: '1. Sele', 'Choo', and 'Hu'.

Home > Analysis > Whole genome analysis

※ 위 분석 페이지는 GiSys Proof Test Bed를 위한 페이지이며, 추후 상용화시에 기능 추가 및 변경이 이루어 질 수 있습니다.

>Human Whole Genome Analysis Procedures

01 02 03 04

Human Whole Genome Analysis Pipeline

The pipeline designed for GiSys-A1

Step	Command
BwaAlign	bwa index;bwa aln
BwaPairing	bwa sampe
Reformatting	samtools view;samtools sort;samtools index
VariantCalling	samtools mpileup
ReportBuilding	make a report

1. Sele

Choo

Hu

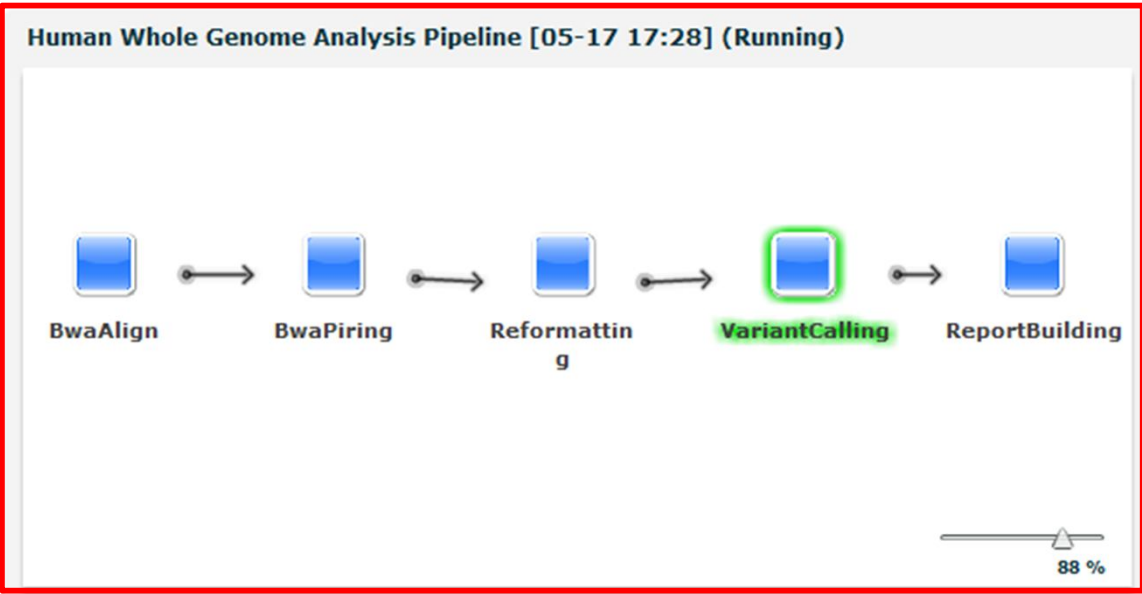
GiSys (Genome Informatics System)



Pipeline List

- [05-17 17:28]Hum: (Selected)
- [05-17 17:07]Hum:
- [05-17 09:23]Hum:
- [05-16 09:50]Hum:
- [05-15 16:08]Hum:
- [05-15 13:36]Hum:
- [05-15 10:34]Hum:
- [05-15 10:25]Hum:
- [05-15 10:05]Hum:
- [05-15 10:05]Hum:
- [05-15 10:02]Hum:
- [05-15 09:34]Hum:
- [05-12 17:15]Hum:
- [05-12 17:08]Hum:
- [05-11 13:31]Hum:

Delete



Name	Description	Executable	Start Date	End Date	Status
BwaAlign	bwa index;bw	/Module/BwaA	2012-05-17 17	2012-05-17 17	Completion
BwaPairing	bwa sampe	/Module/BwaP	2012-05-17 17	2012-05-17 17	Completion
Reformatting	samtools view	/Module/Refor	2012-05-17 17	2012-05-17 17	Completion

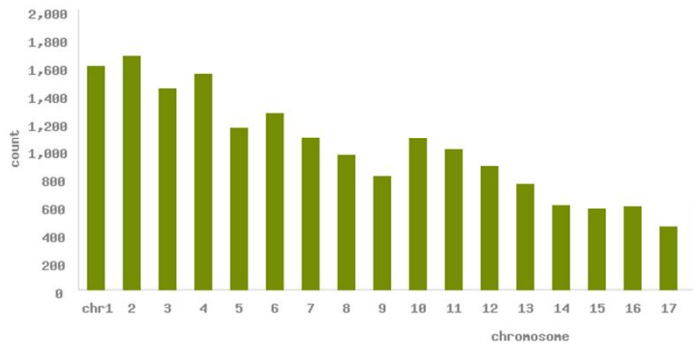
GiSys (Genome Informatics System)

> SNV (Single Nucleotide Variation)

- 단일염기서열변이 : 하나의 서열 또는 종내 소수의 집단에서 나타나는 단일염기의 차이로, 시퀀싱 데이터에서 나타나는 표준 염기서열과의 차이.

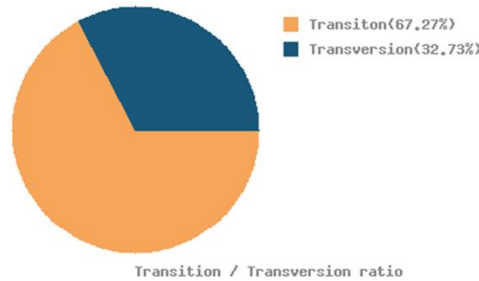
1. SNV count

- 단일염기서열변이(SNV)의 개수
- 정렬된 read 서열을 이용하여, reference 서열과 샘플 염기서열의 비교를 통해 샘플 특이적인 다형적 유전자형을 확인한다. 다형적 유전자형은 높은 신뢰도의 SNV 데이터를 생성하기 위해 필터링 과정을 거친다. 게놈의 coding 지역의 SNV의 발생에 의해 아미노산이 치환 및 단백질의 구조 및 기능의 변화를 일으킬 수 있다.



2. Transition / Transversion

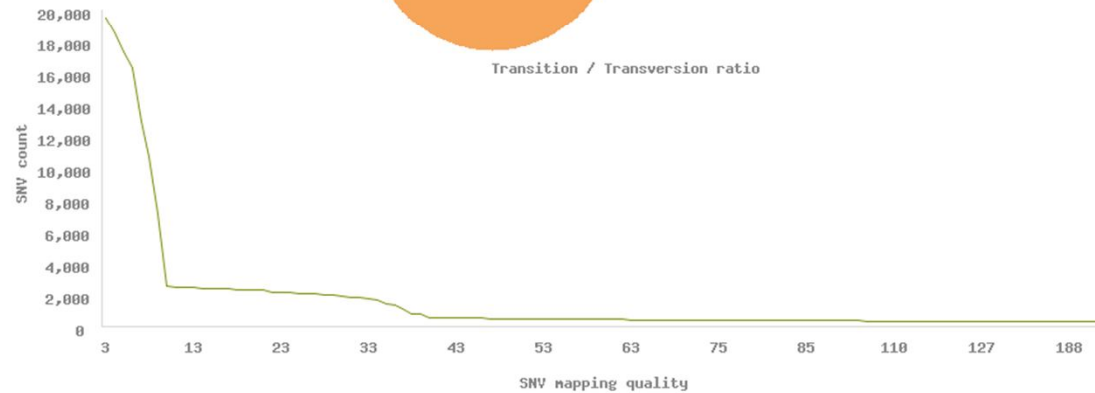
- 유사성 염기치환 (Transition) : 퓨린이 다른 퓨린에 의해 또는 피리미딘이 다른 피리미딘에 의해 교체된 염기치환
- 교차형 염기치환 (Transversion) : 퓨린에서 피리미딘으로 대체되고, 피리미딘이 퓨린으로 대체되는 염기 치환
- 일반적으로 모든 계통상의 모든 DNA 서열에 있어 Transition(T<->C, A<->G)의 빈도가 Transversion (T<->A, T<->G, C<->A, C<->G)의 빈도보다 높다 (Brown et al. 1982; Gojobori et al. 1982; Curtis and Clegg 1984; Wakeley 1994, 1996). Transition / Transversion 비율은 DNA 서열 진화의 일반적 특성이라고 알려져 있다 (Wakeley 1996). 또한, Transition / Transversion 비율은 normal 샘플과 cancer 샘플에서 차이를 보이는 특징이 있다.



Transition count	Transversion count	Transition / Transversion ratio
13,270	6,456	2.055

3. SNV mapping quality

- 정렬된 read 서열 정보를 이용하여 측정된 SNV mapping quality의 누적 그래프로, 특정 m.



SUMMARY

- 2006, PGP
- 2008, Korean Genome Project
- 2010, KPGP
- 2011, KPGP–20
- 2012, OPENKPGP – TOTAL 38 genomes
- **GiSys** (Genome Informatics System)

TBI and PGI

- **TBI** (Theragen BiO Institute) of **TheragenEtex**
 - For profit commercial entity
- **PGI** (Personal Genomics Institute) of **GRF** (Genome Research Foundation)
 - Non-profit, Government managed org.



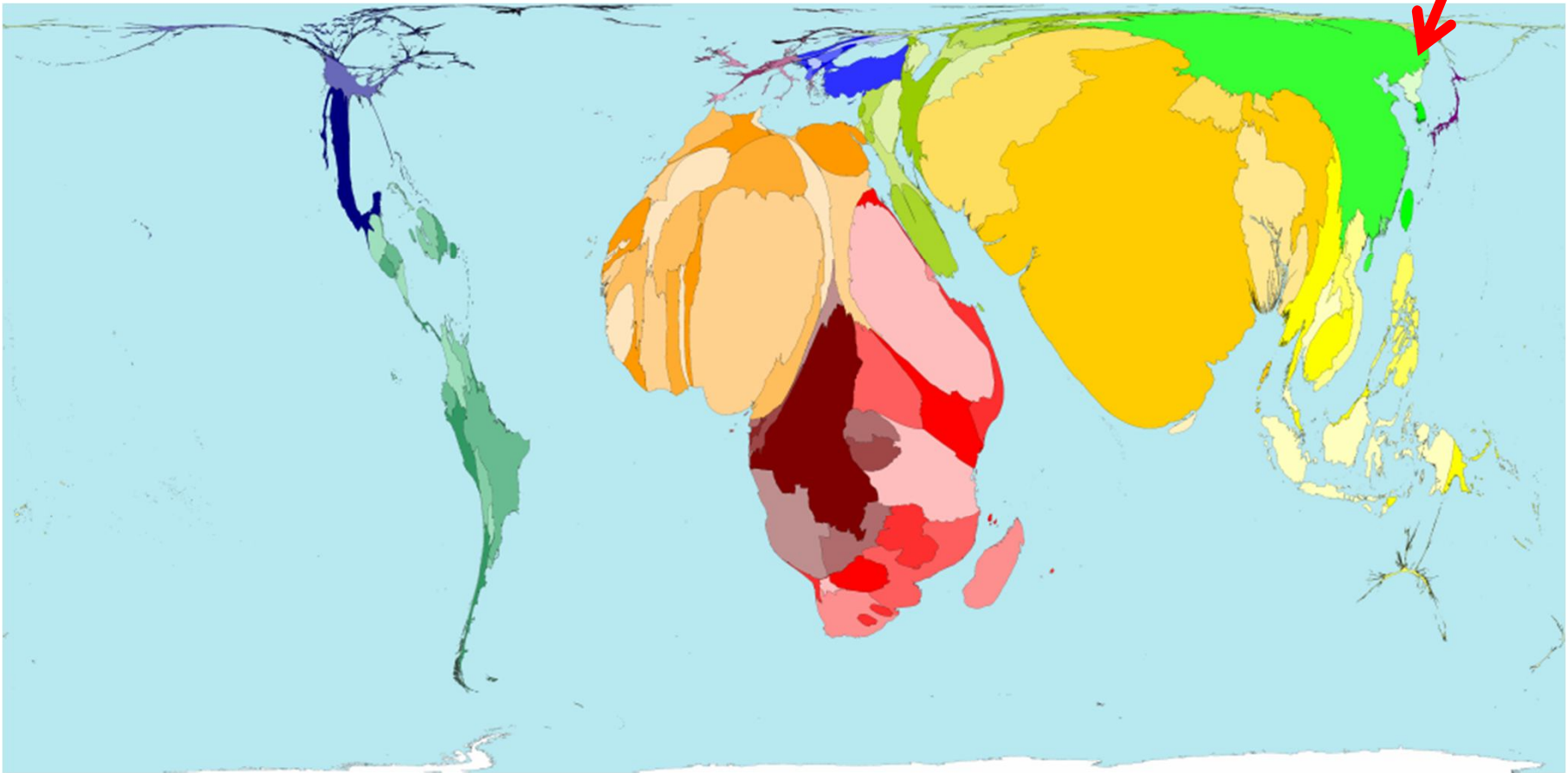
TBI Projects

- HelloGene/ HelloGenome → Personal gene information service
- Tiger (Theragen)
- Whale (with KORDI)
- Gastric Cancer (NCI, Korea)
- Lung Cancer
- Spleen Cancer
- Pancreatic Cancer
- Bladder Cancer
- Korean Horse, Korean Cow, Dog, Monkey, Birds,
- Soybean

Totalomics

- Totalomics: Sequencing and Bioinformatics service by TheragenEtex.
- <http://totalomics.com>
- <http://totalomics.kr>

What map is this?



유아사망률