# NGS를 이용한
# 게놈분석

Jong Bhak
박종화
게놈연구재단

2012. 05.02

# 결론 1.

Let’s do more sequencing

시퀀싱 더 하자!

By Jong Bhak

# 결론 2.

Sequencing 7 billion people
on Earth
as cheaply as possible

70억명 해독

By Jong Bhak

# 감사의 말

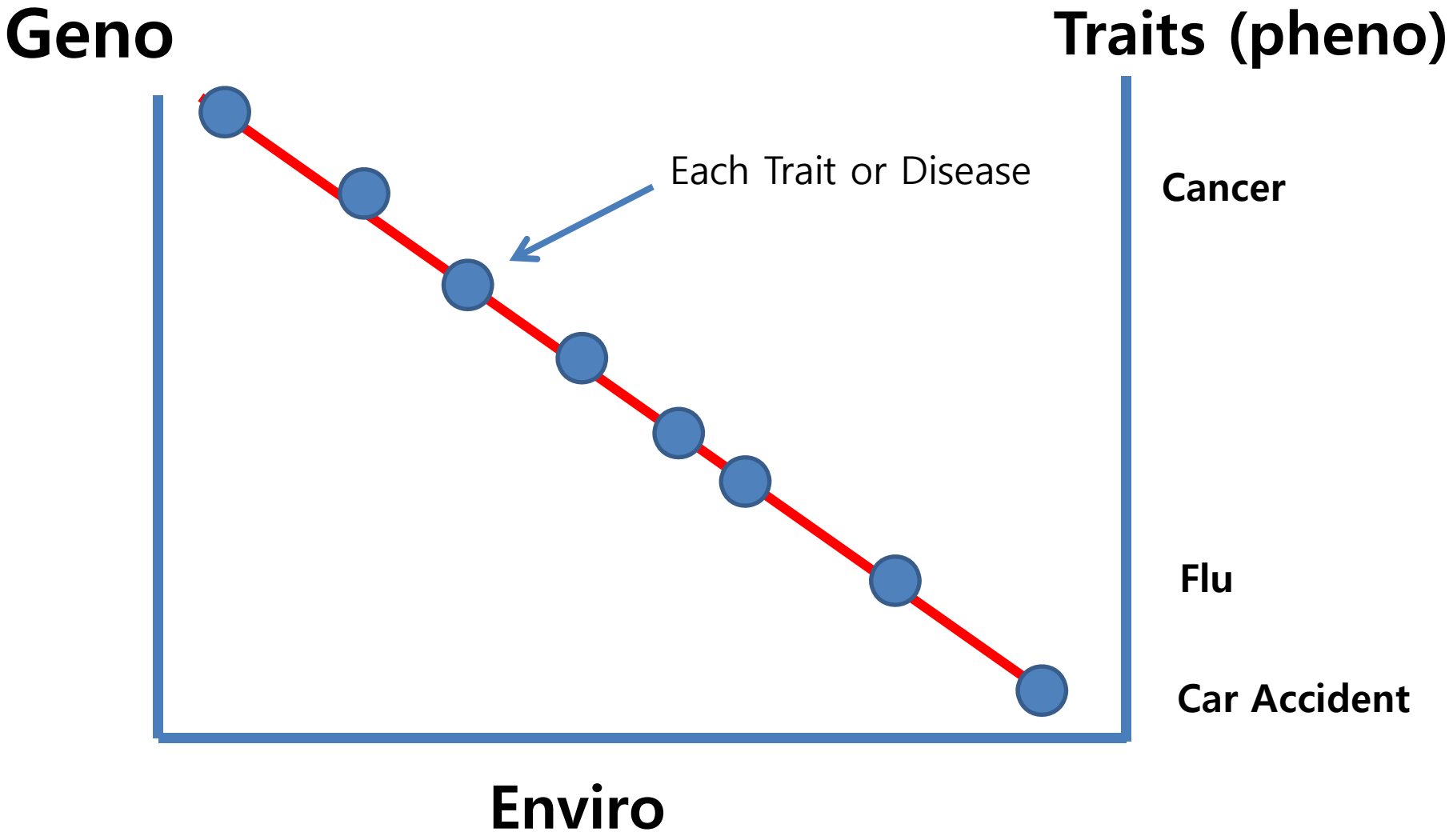
- 연구자체에 열정을 가지고 정확한 데이터를 내는 많은 국내외 과학자들
- 과학기술인을 지원하는 국민들
- 게놈관련 상품을 사주는 고객들
- 성신여대 와 김상태교수님
- 테라젠 및 게놈재단의 많은 동료들
- 재정적 지원(기부)를 해준 테라젠 (고진업대표)

# 게놈?

# Gene, Environe, Phene (유전자, 환경자, 형질자)

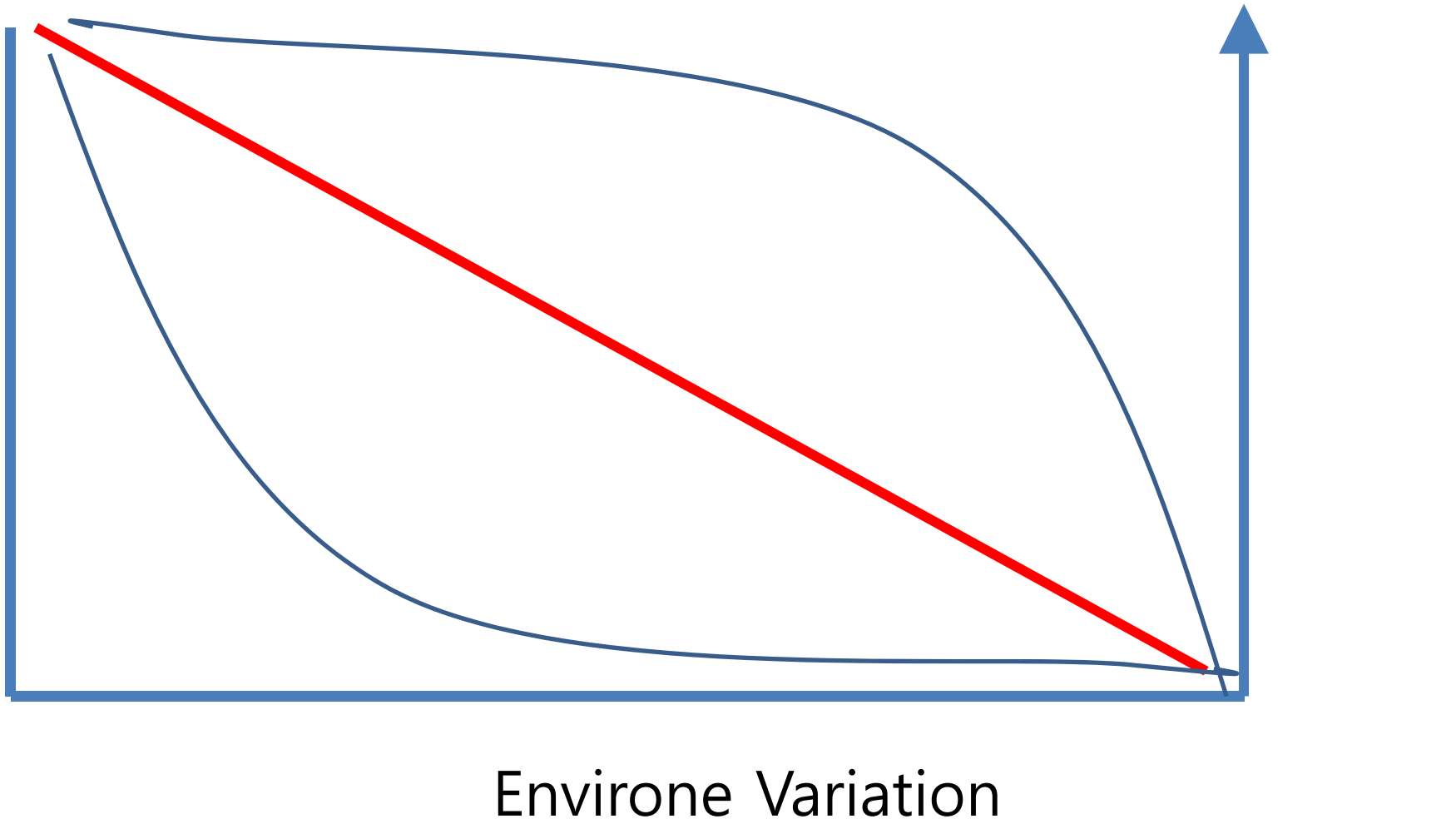
- Gene: single genetic factor (genetic atom)
- Environe: single environmental factor (temp.)
- Phene: single phenotype (trait: hair color)
- Genome, Envirome, Phenome (Traitome)

# Geno + Enviro = Pheno (GEP graph)



# Single Gene . Environe . Phene Variation

- Gene Variation



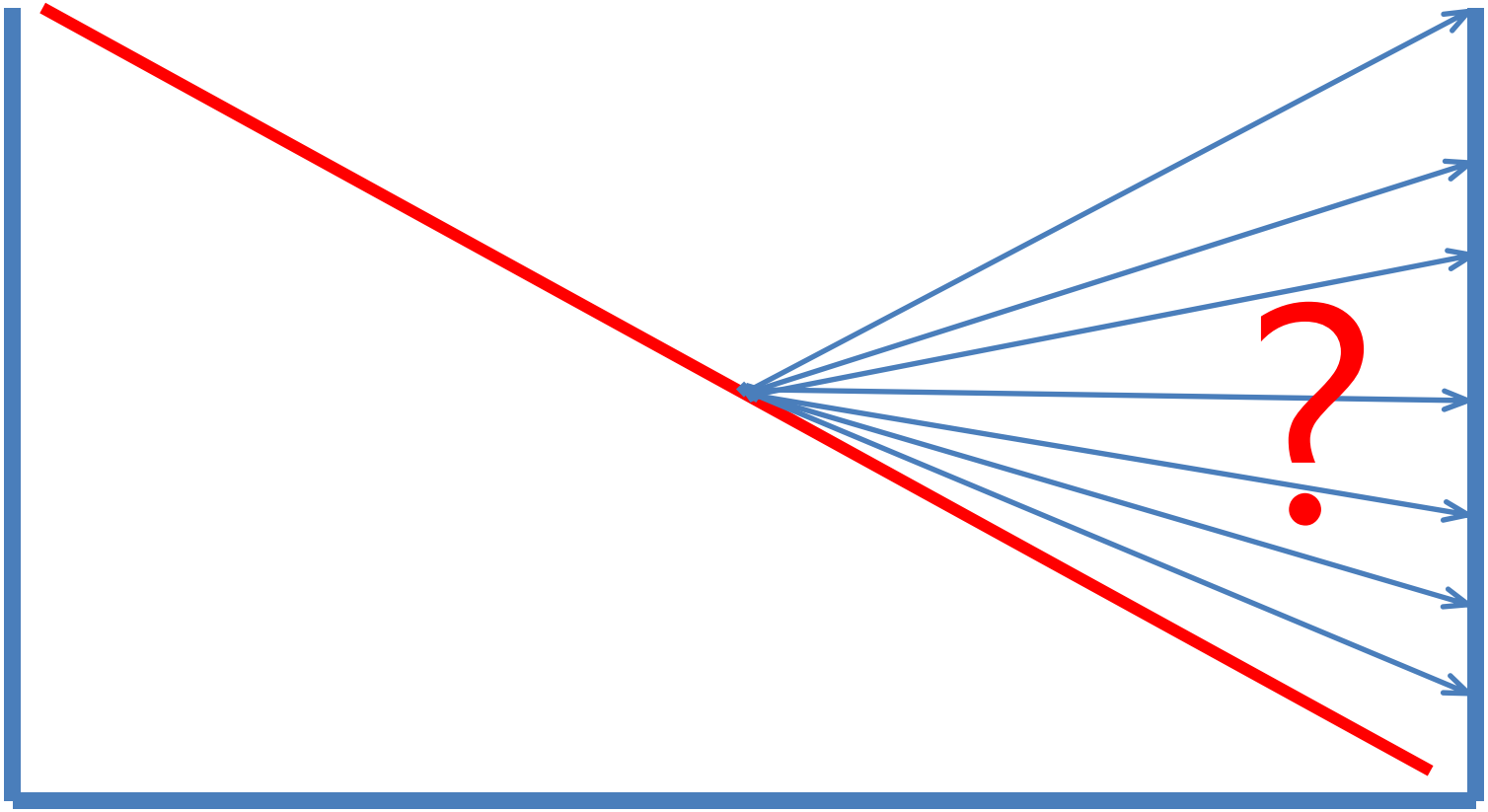
# Genome Envirome and Phenome

- Genome = gene types + their variome
- Envirome = environe types + their variome
- Phenome = phene types + their variome

# GenoEnviroPheno Unpredictability Graph

- **Genome**

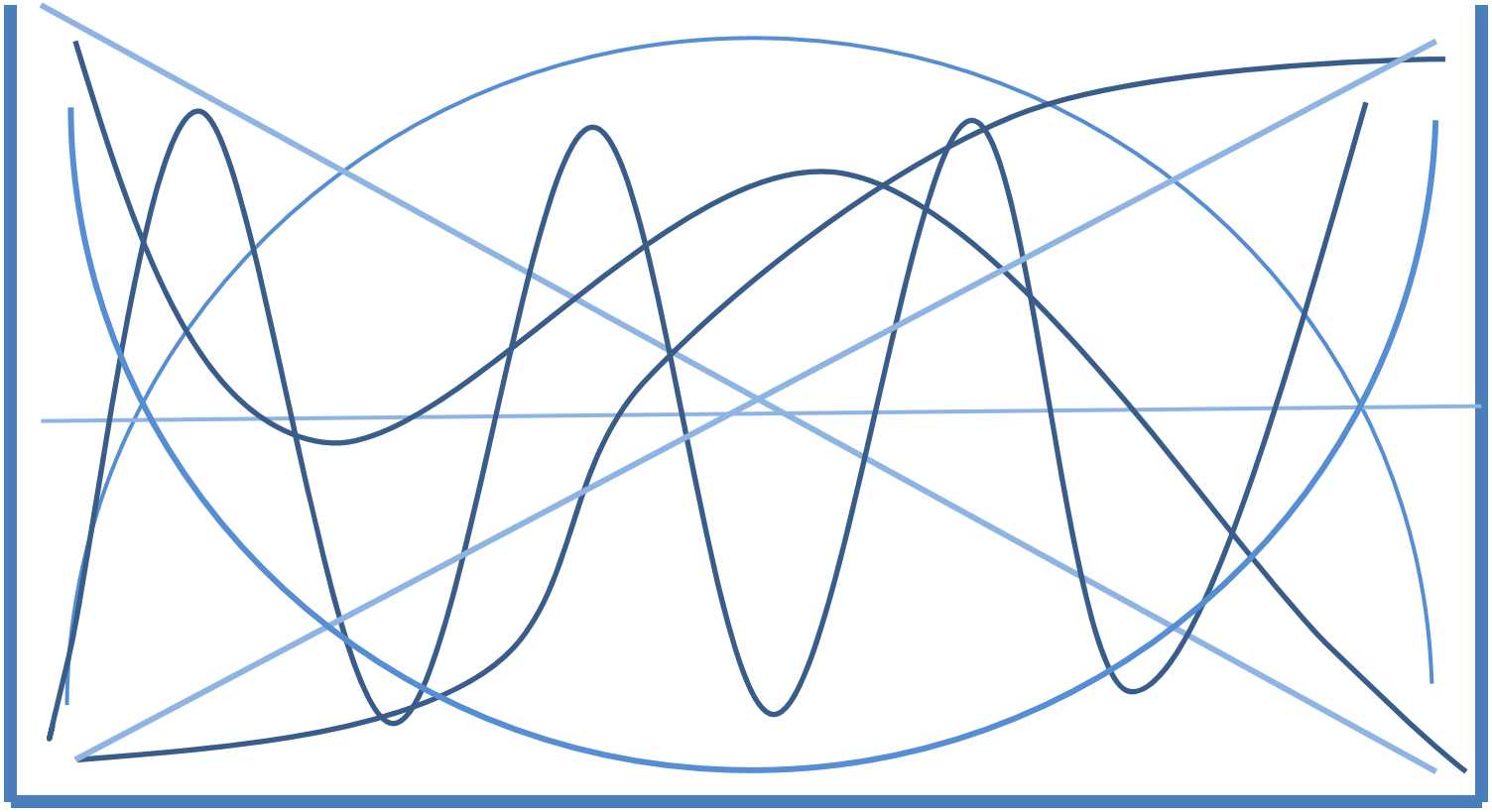
**Phenome**



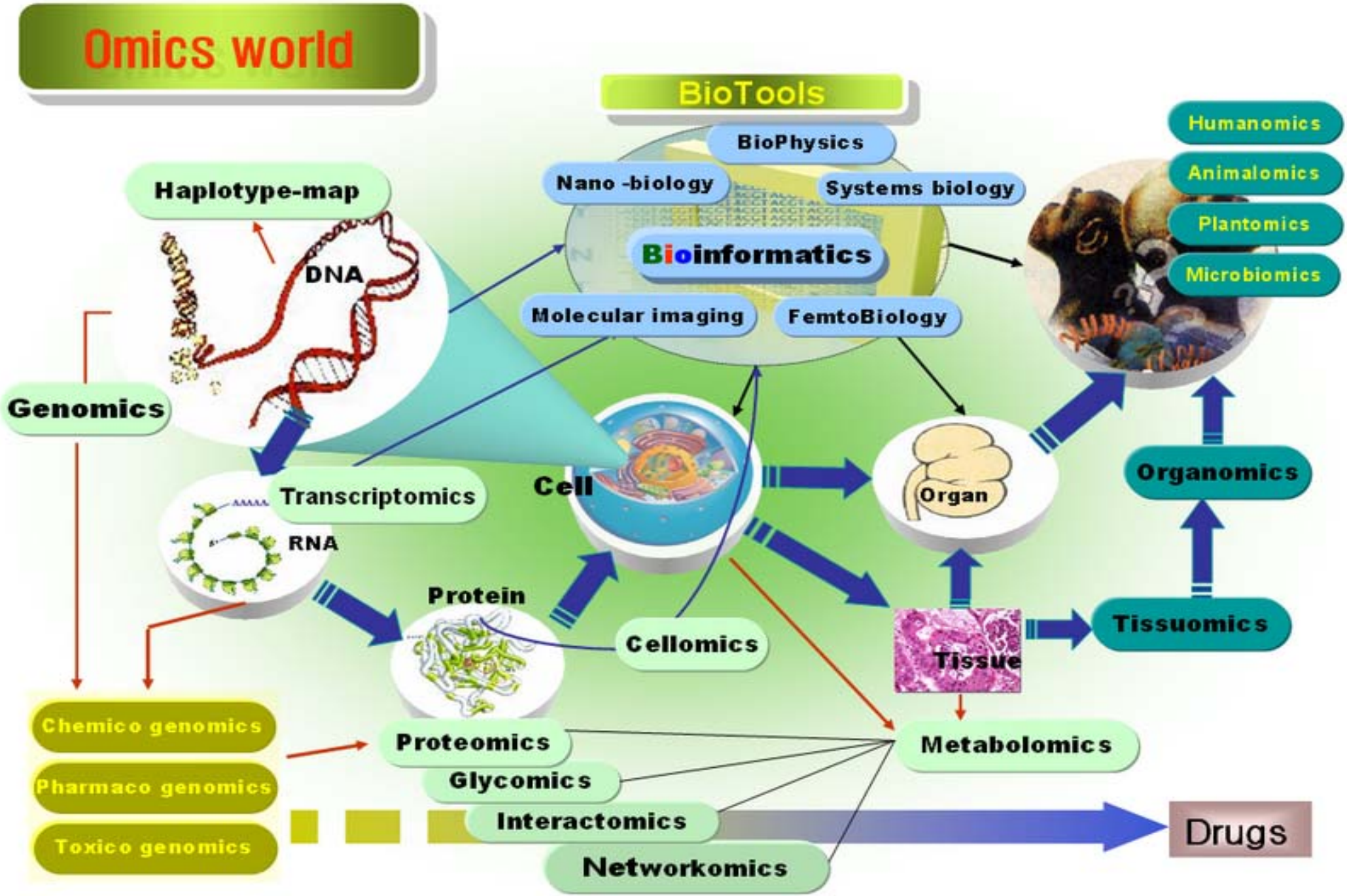

**Envirome**

# Genecomplex ←→ Phenecomplex

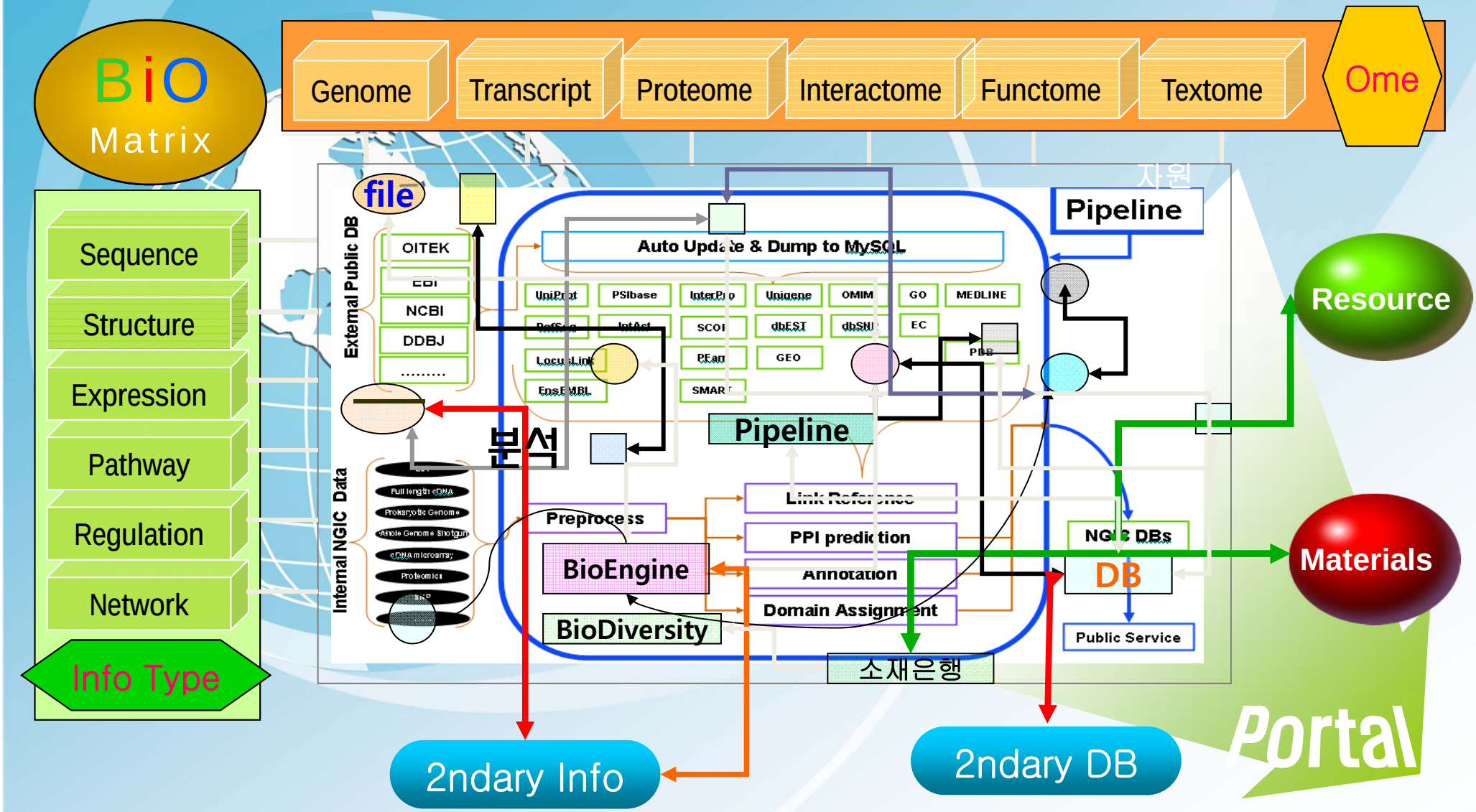
- GeneComplex



PheneComplex


Omics world
BioTools
BioPhysics
Nano -biology
Systems biology
Bioinformatics
Molecular imaging
FemtoBiology
Humanomics
Animalomics
Plantomics
Microbiomics
Haplotype-map
DNA
Genomics
Transcriptomics
RNA
Cell
Organ
Organomics
Protein
Cellomics
Tissue
Tissuomics
Chemico genomics
Pharmaco genomics
Toxico genomics
Proteomics
Glycomics
Interactomics
Networkomics
Metabolomics
Drugs

# Biomatrix: Putting structure in omics



# 게놈학이란?

- Genomics is the bioinformatic study of genes of individual organisms, populations, and species.

- http://genomics.org

# DNA 해독 and DNA 타이핑

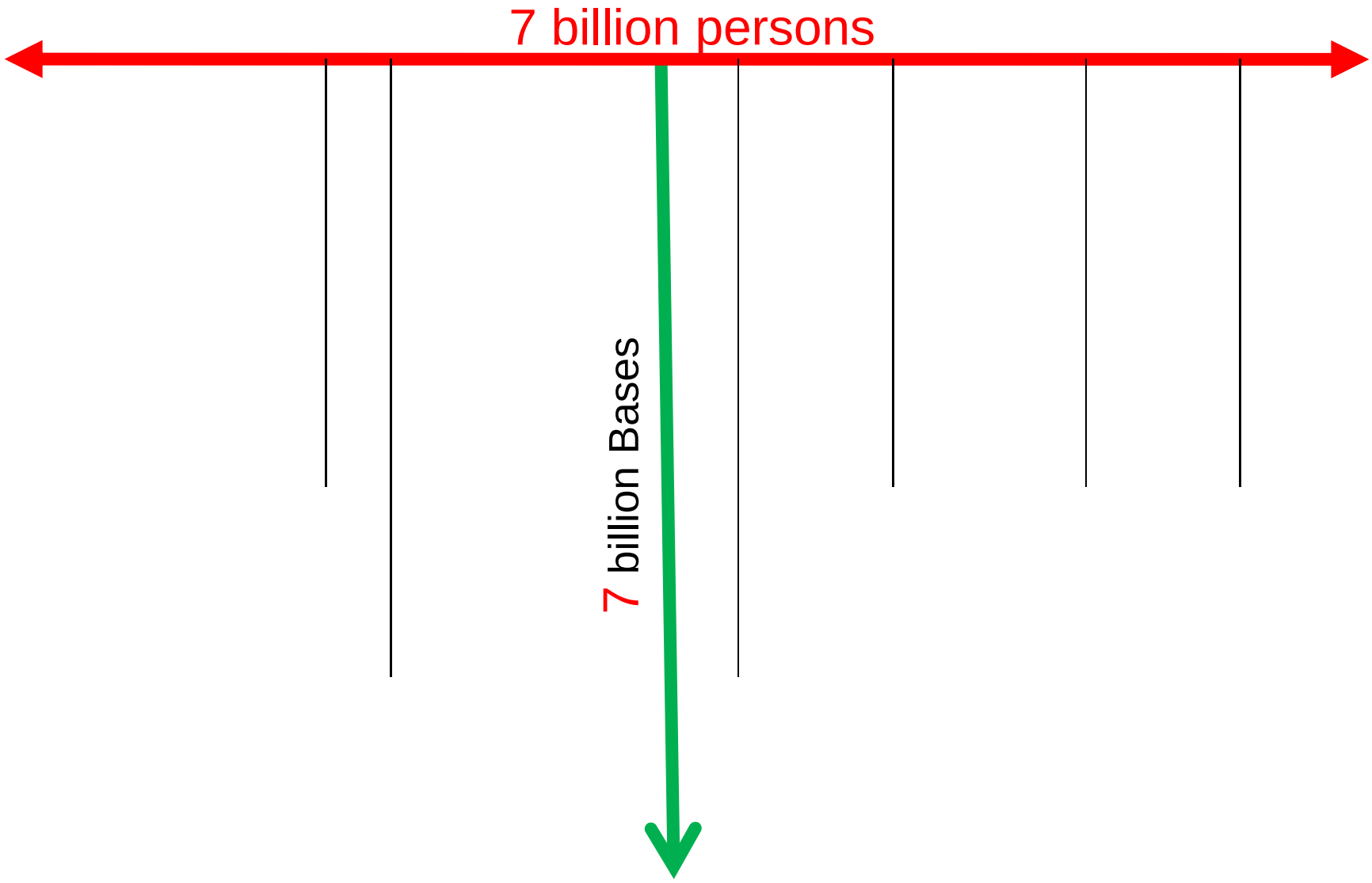
- 해독 ➔ 궁극적

- 타이핑 ➔ cheap and efficient

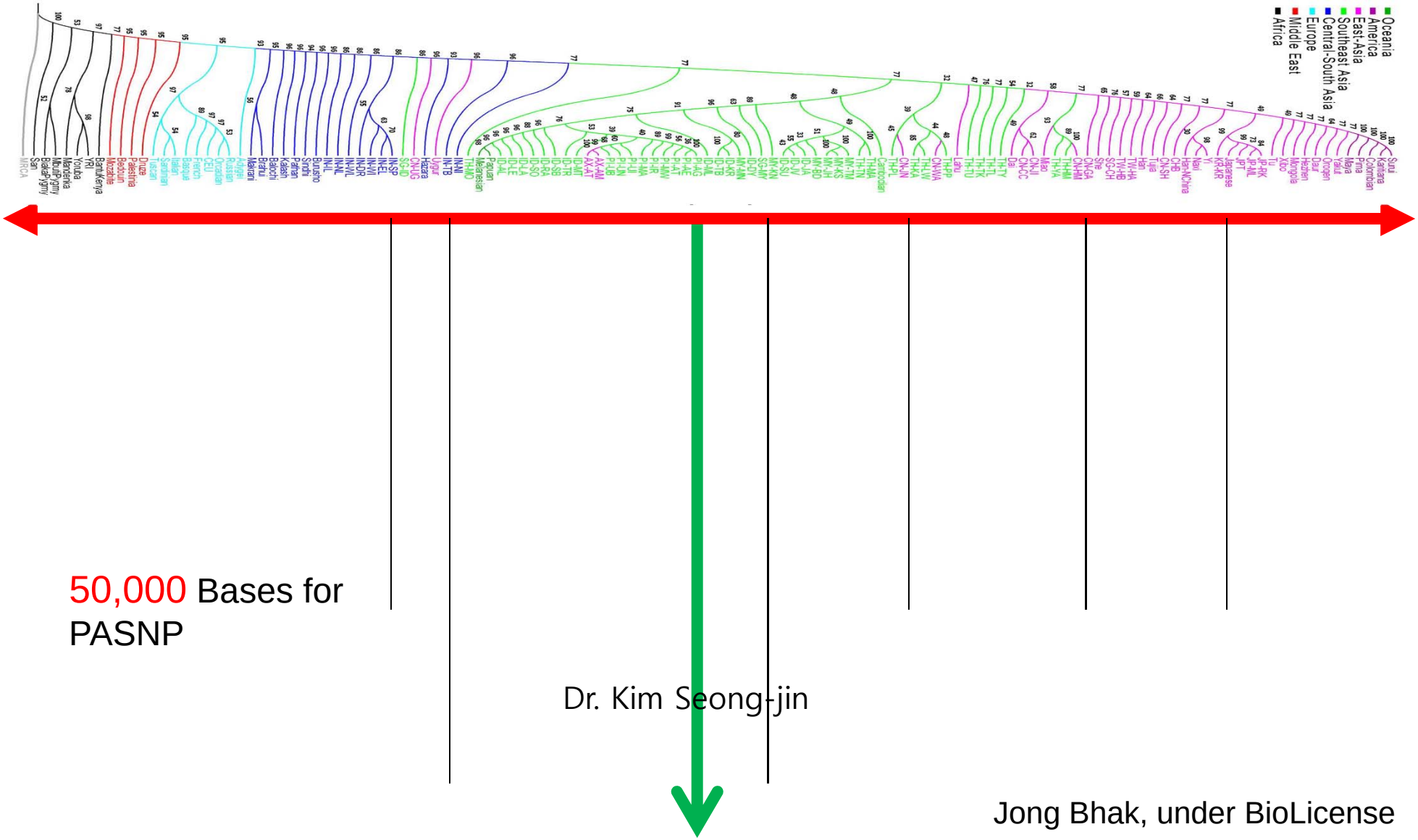

# 차세대 해독기

- 일루미나
- 라이프테크 (구 AB)
- 로슈
- IBS
- CompleteGenomics
- PacBio
- Helicos
- NanoPore
- Genia

# Genomic T :



# Genome Diversity and Genome Variation



# nsSNVs in proteins

# The first Korean Genome Sequence




## 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

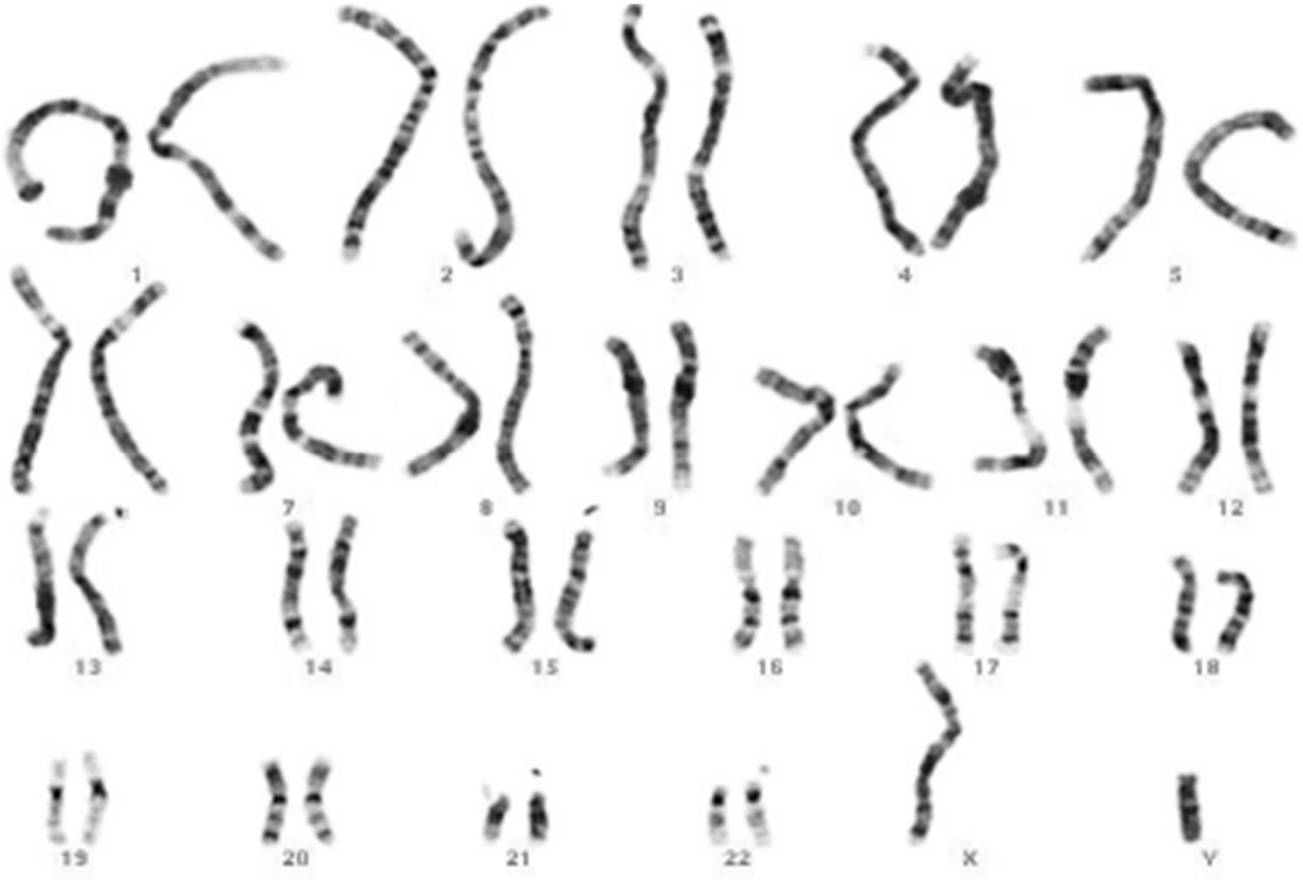
Data publicized: 2008. Dec. ftp://bioftp.org

# The first Korean Genome (SJK)

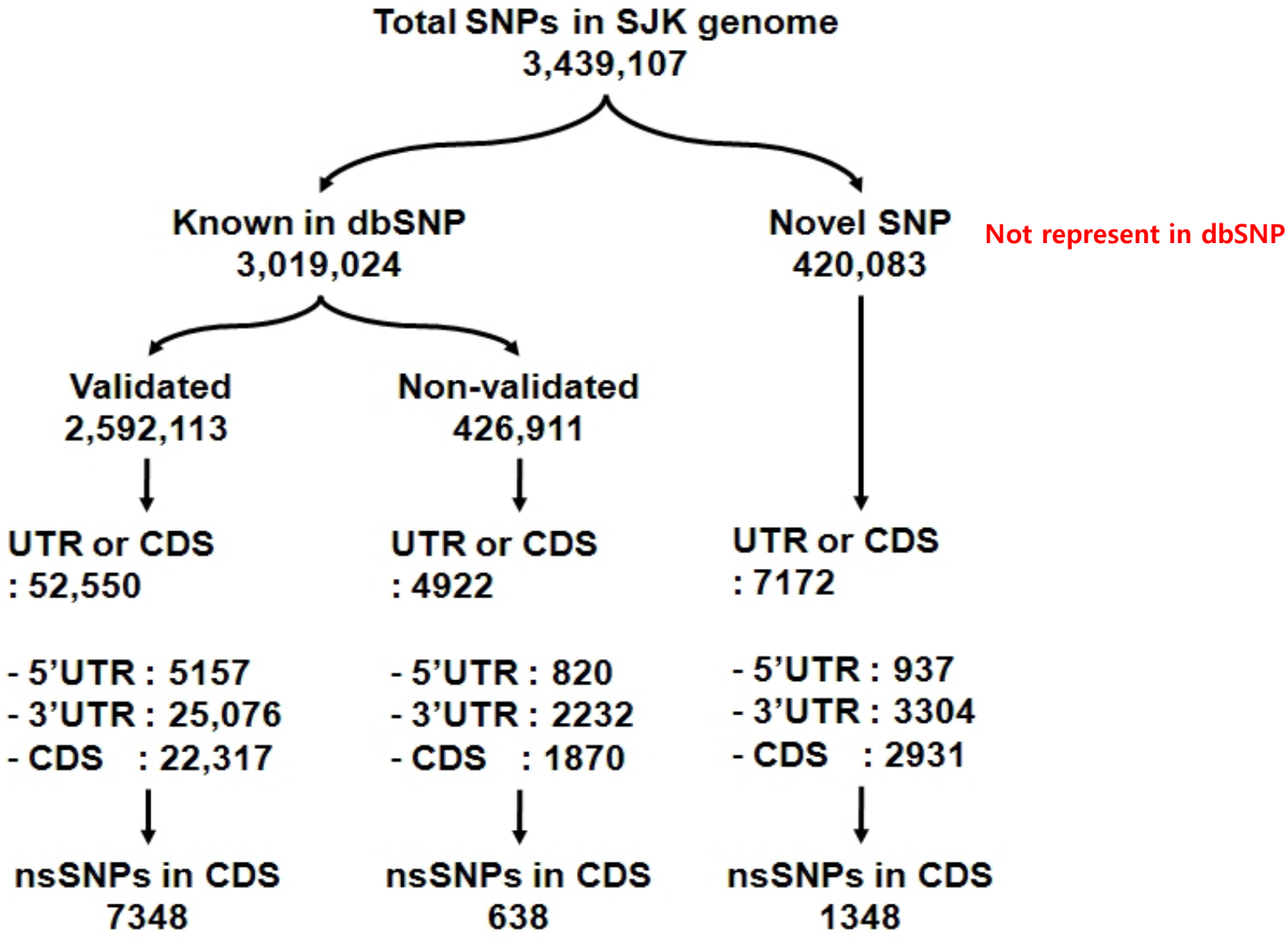
- First analyzed by Gacheon medical school LCDI and KOBIC, KRIBB in 2008 (Joint effort among LCDI, KOBIC, and 국가참조표준센터)
- First annotated and made public on 4th Dec. 2008 (through web and ftp)
- SNP, CNV, indels were analysed
- Automated phenotypic association study was done
- Non-syn. Analysis
- Phylogenetic study of mtDNA, Y Chr And autosomes showed Korean relationship to Chinese and Japanese.
- First intra-Asian genome comparison (Chinese and Korean)
- Analyzed at: 7.8, 17.3, 23.5 and 28 x folds
- By Jan. 23.5 fold sequenced and analyzed
- Openfreely Available from: http://koreagenome.org

# The Karyogram of the donor DNA

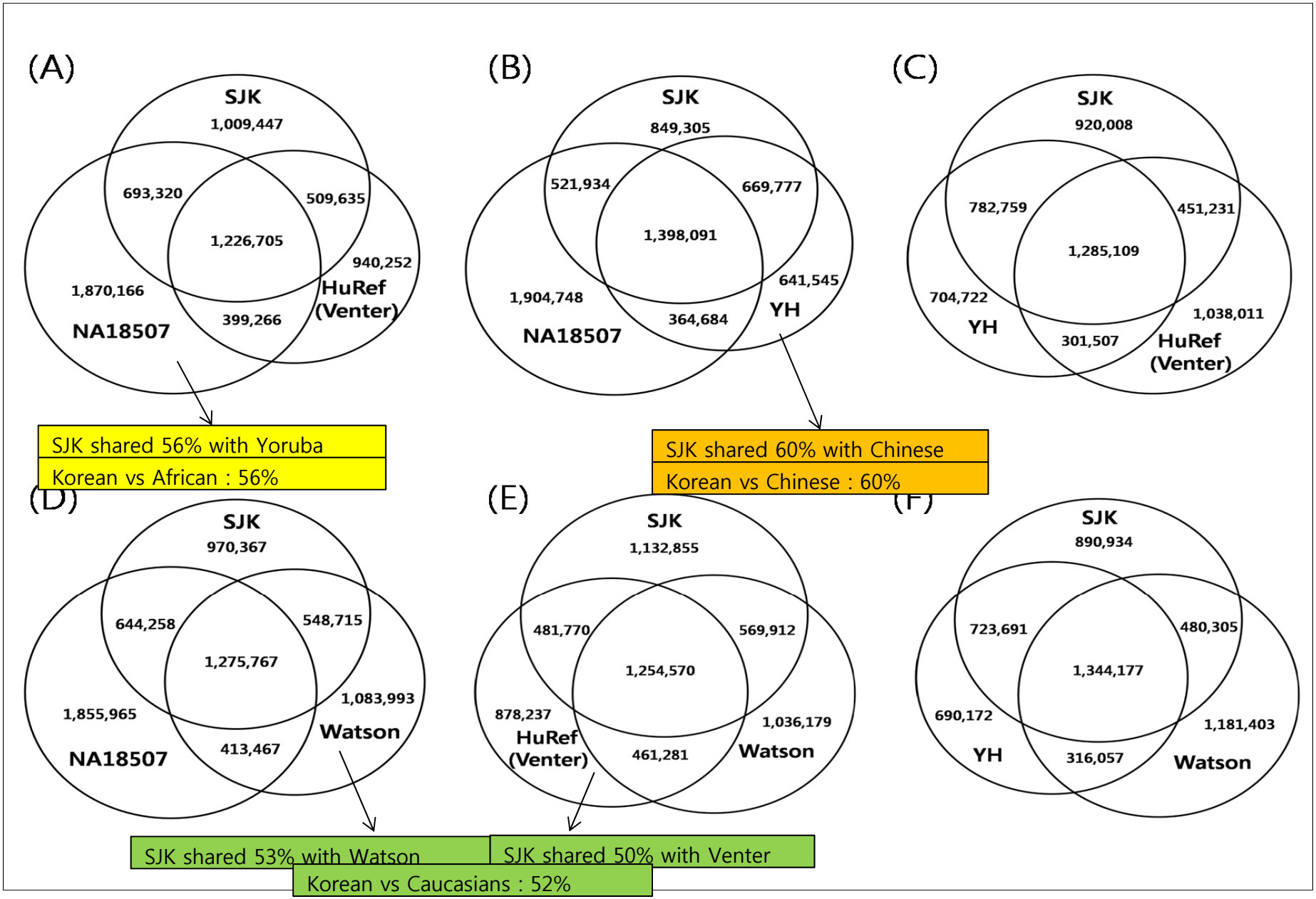
**No obvious chromosomal abnormalities!**



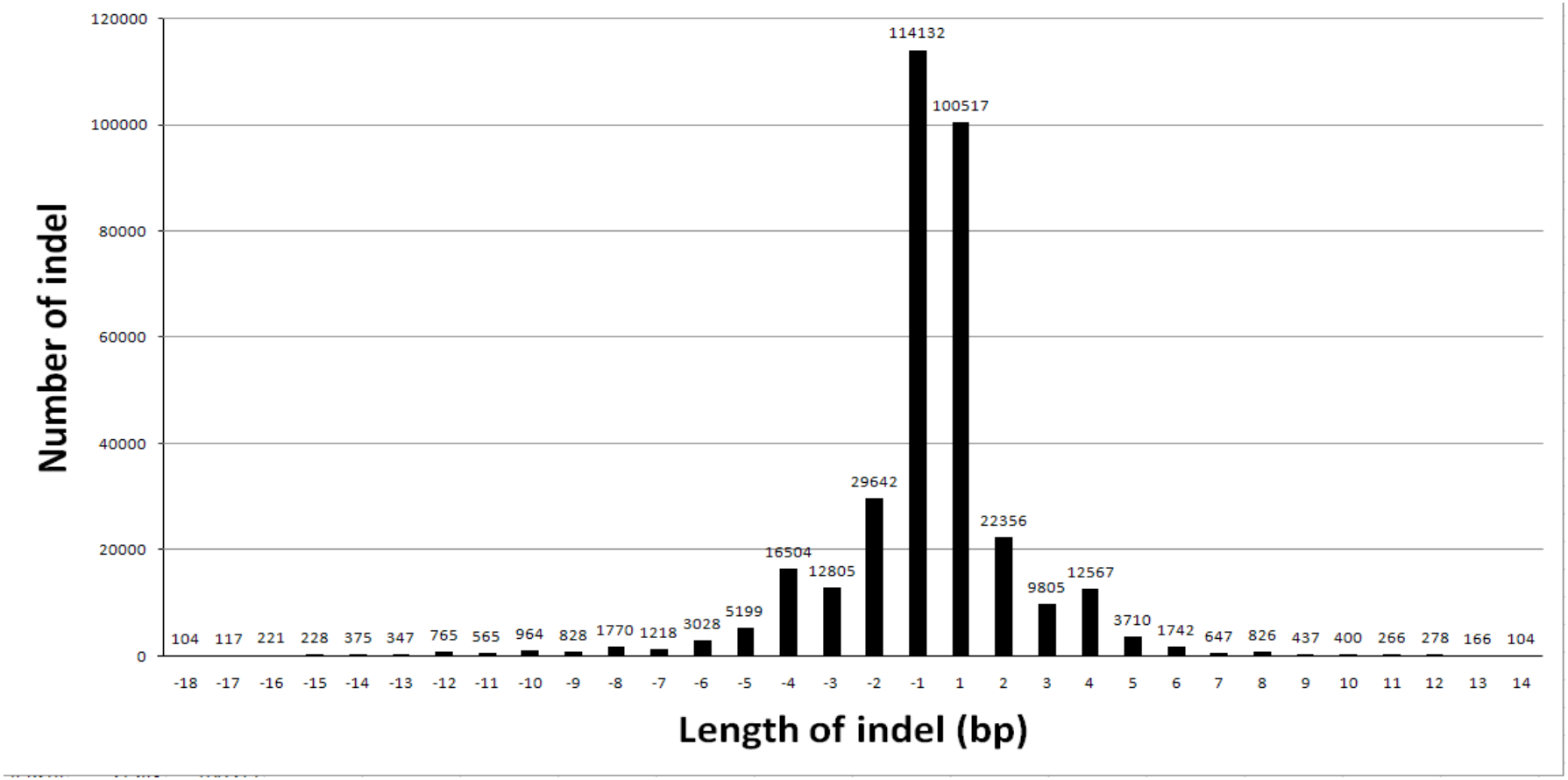
# Classification and number of intra-genic SNPs



# Comparison of individual SNPs



# Size distribution and classification of short indels found in SJK

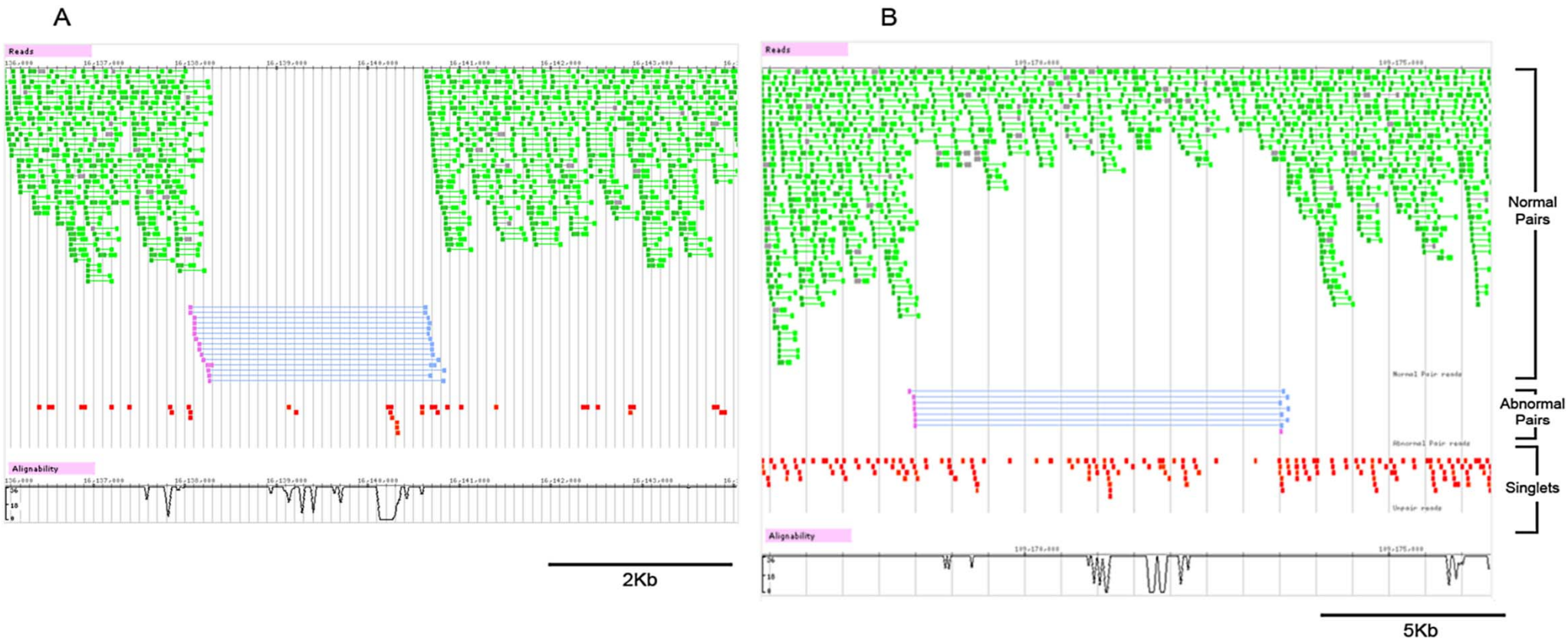



Using MAQ, we identified 342,965 short indels
→ We found that only **247** (0.1%) were validated,**113,287** (33.0%) non-validated, and **229,431** (66.9%) indels were not found in dbSNP

# Indels in SJK genic regions

| Index | Indel | | | |
|---|---|---|---|---|
| | Indel number | Homozygous | Heterozygous | Gene number |
| 5'UTR | 27 | 9 | 18 | 26 |
| CDS | 49 | 16 | 33 | 40 |
| 3'UTR | 319 | 114 | 205 | 247 |
| Intron | 127,516 | 45,430 | 82,086 | 12,421 |
| Total | 127,911 | 45,569 | 82,342 | 12,734 |

# Homo- and heterozygous deletions in SJK genome



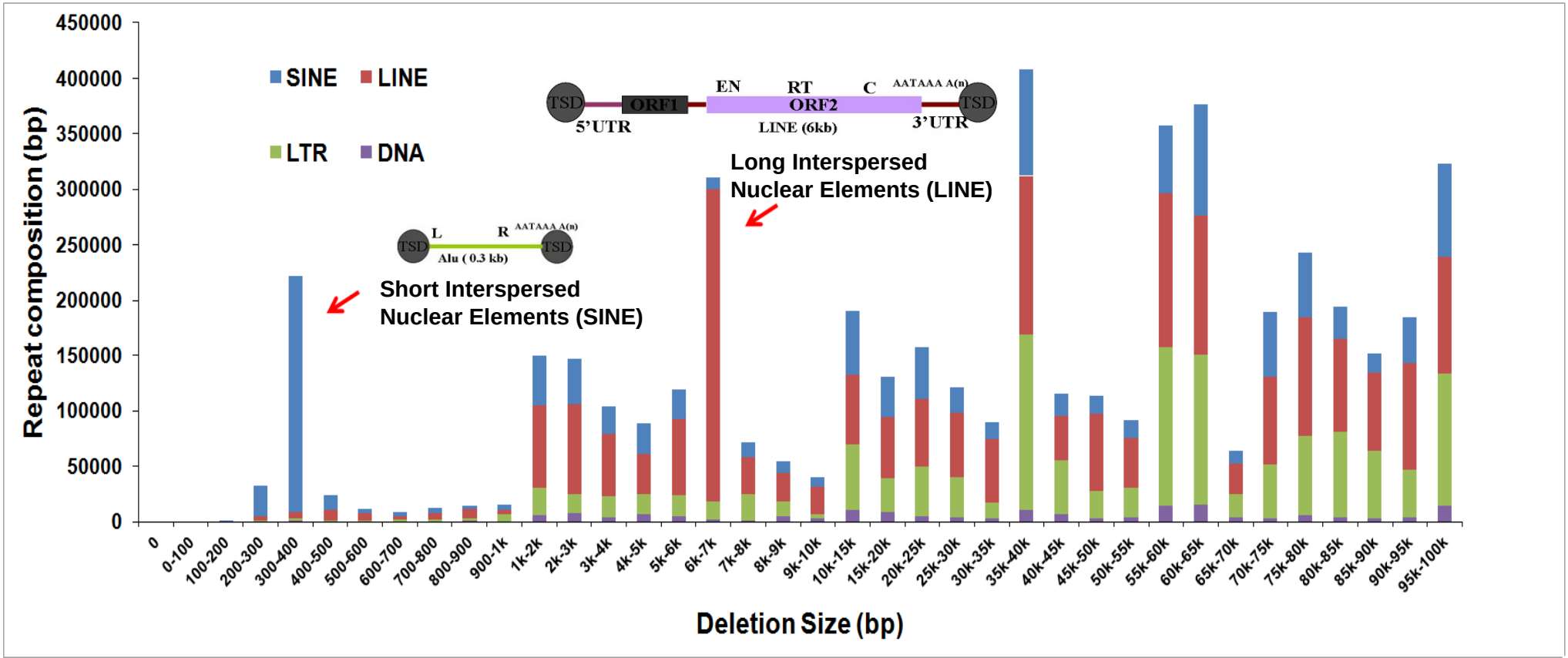

**(A) Homozygous 2.3 kb genomic deletion and (B) Heterozygous 5 kb genomic deletion.**

# Detection and identification of structural variants

- **We found structural variants by using paired-end reads.**
  1. 2920 deletions (100bp ~ 100kb)
  2. 415 inversions (100bp ~ 100kb)
  3. 963 insertions (175bp ~ 250bp)

- **We found deletion SVs in 21 coding genes.**
  → All heterozygous deletions

# Repeat composition in SJK deletion variants



# Soybean 유전체 분석

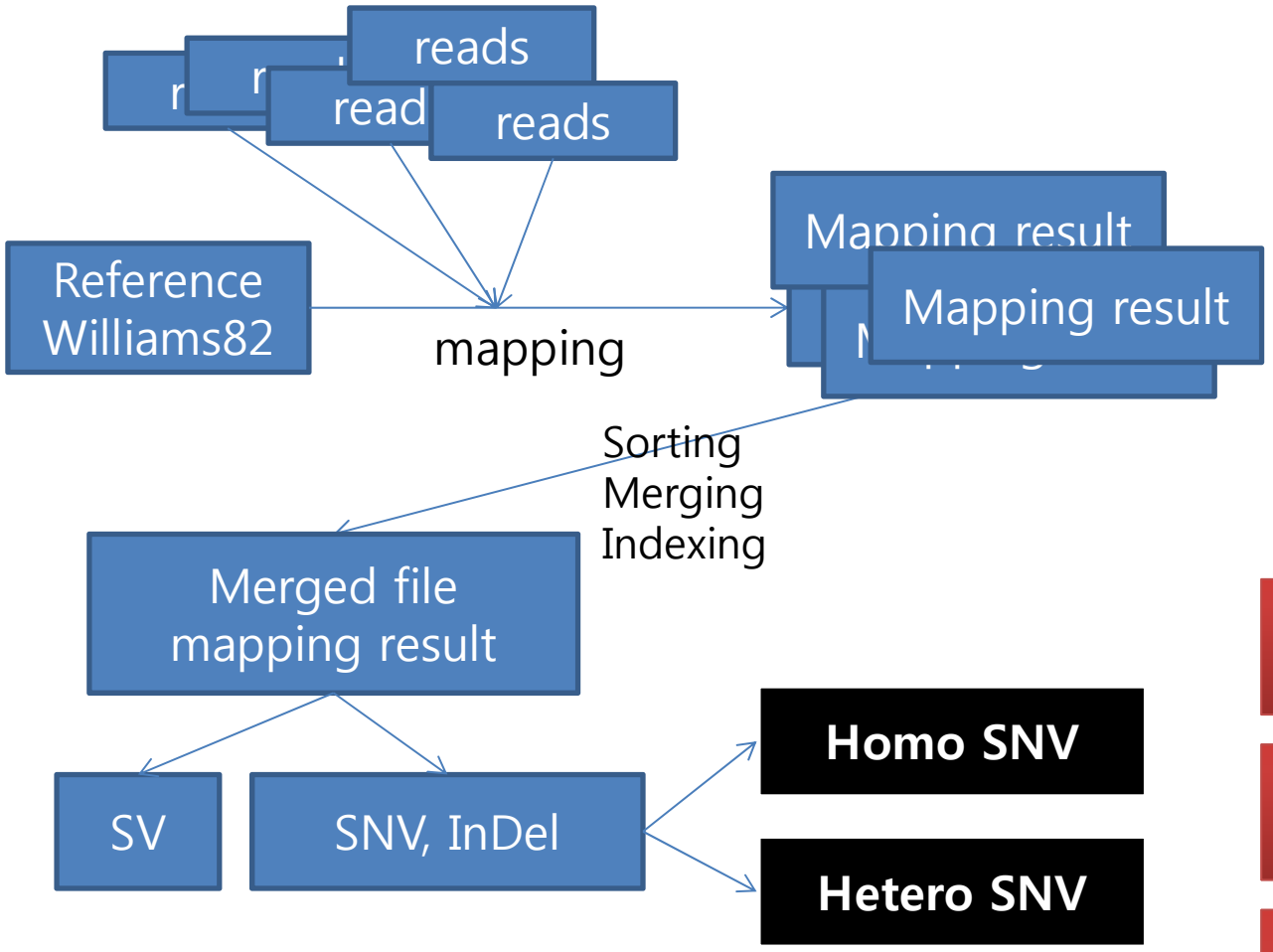
# Overview

- Sequencing: *G.max*　품종 1, 품종 2
  - Illumina GAIIx

- Reference : G.max Williams 82 (Glyma1)
- Tools : bwa, samtools, breakdancer

- 분석항목
  - 1차분석
    - 레퍼런스 맵핑 통계
  - 2차분석
    - SNP
    - Small Indel
    - SV
  - 3차분석
    - Comparison
    - Pattern analysis
    - Pathway mapping

# Analysis flow




**Comparison**
품종 1, 2, soja

SMV resistance

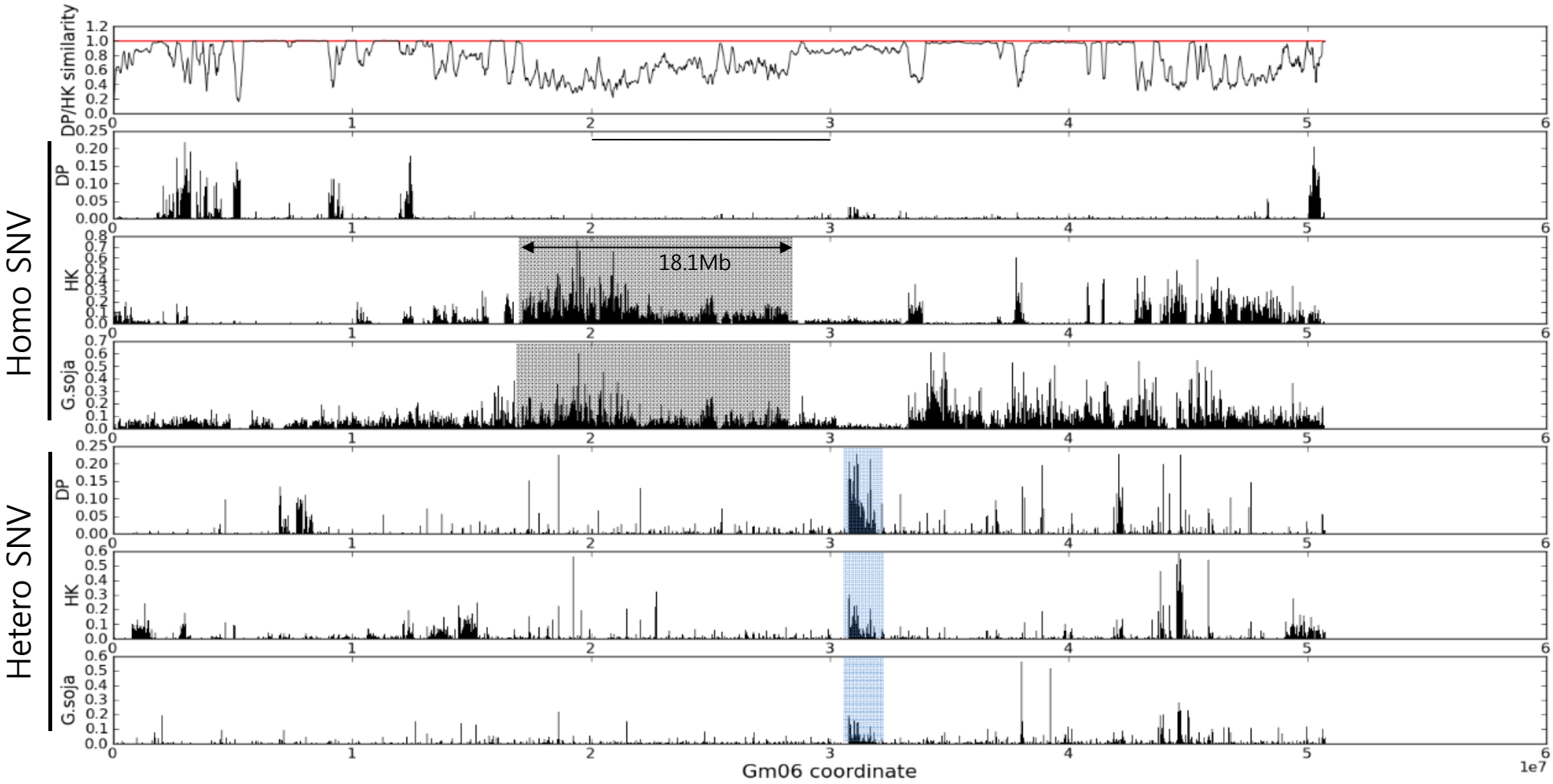
nsSNV

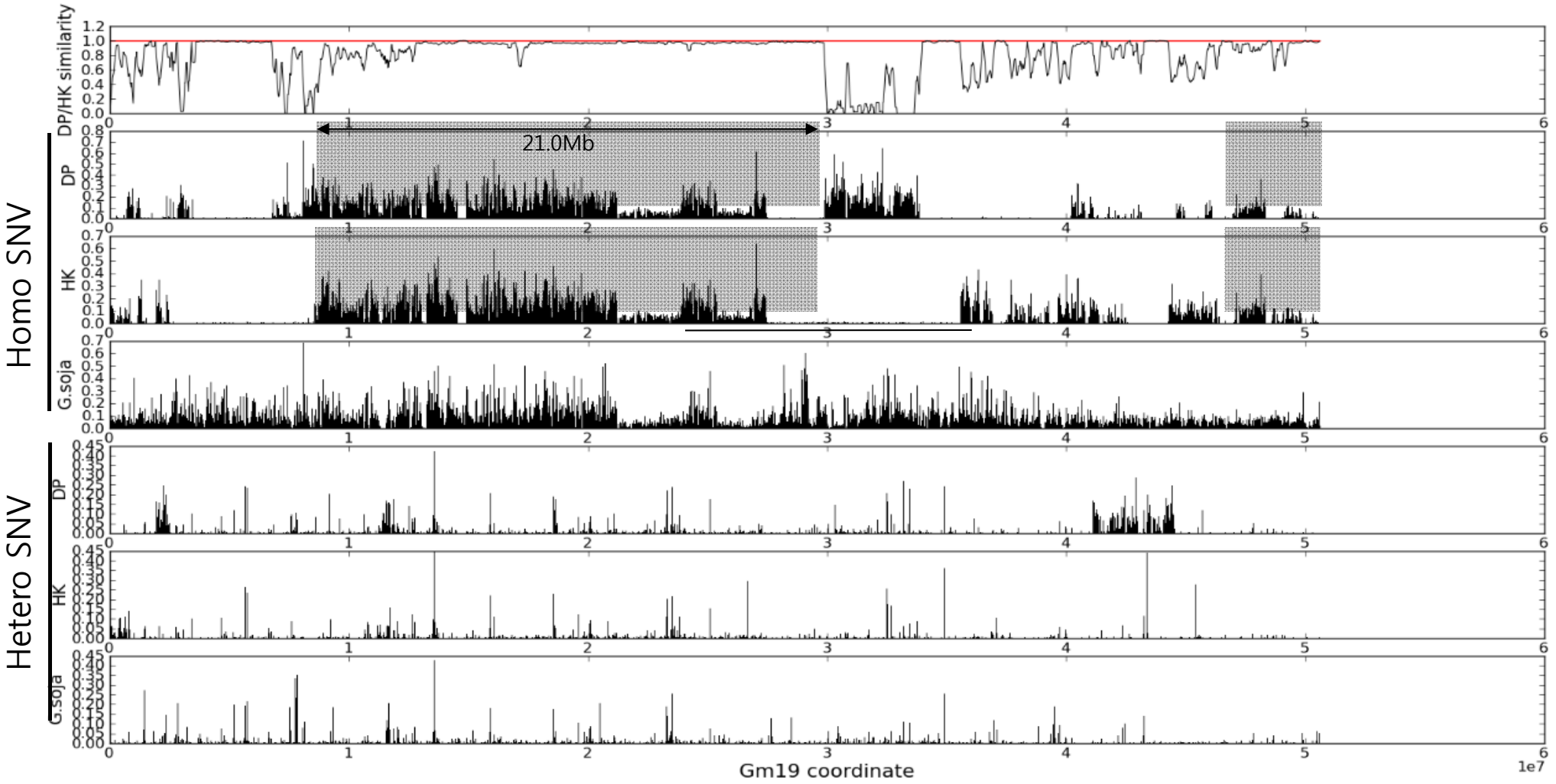
SNV
density

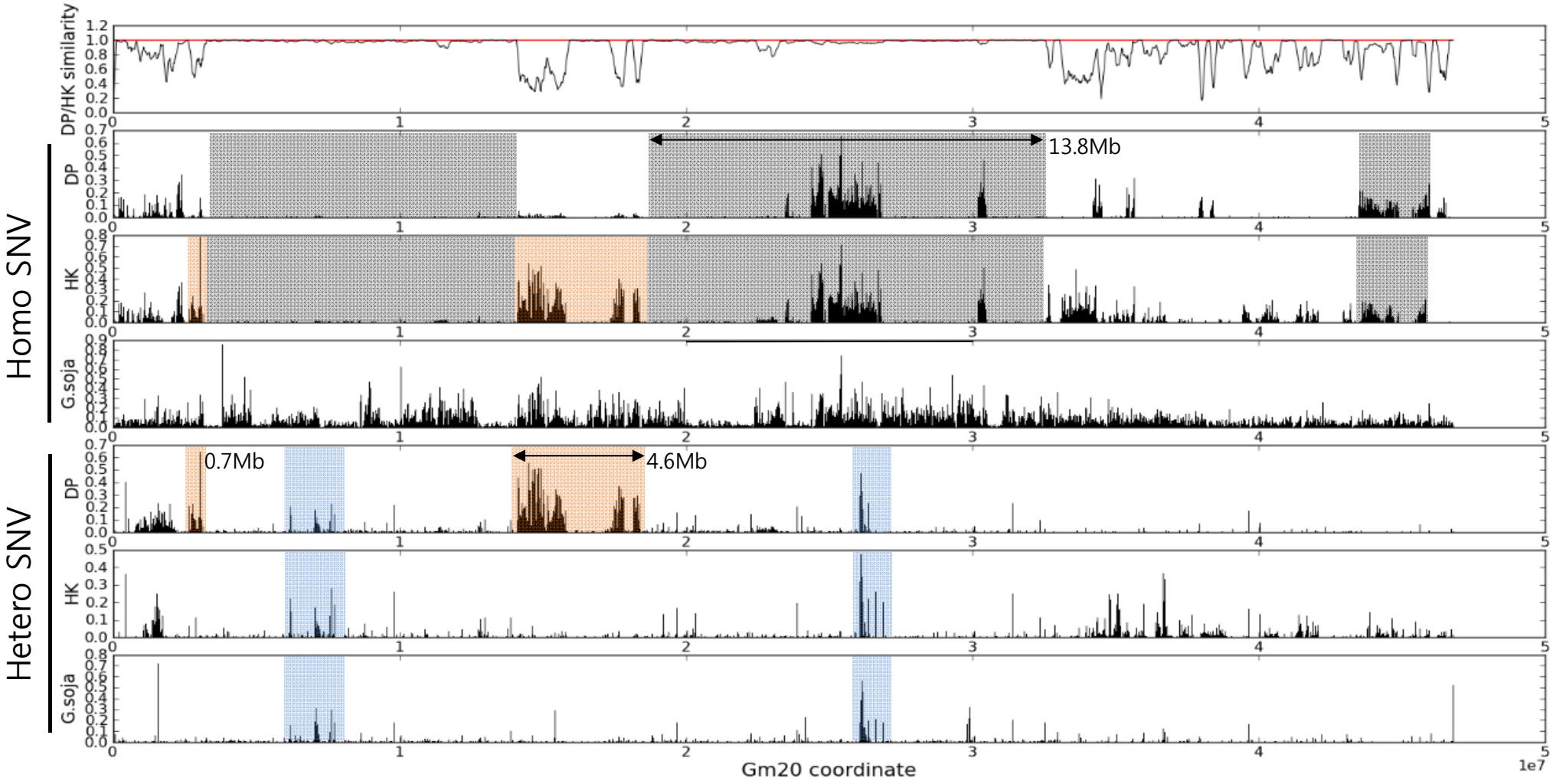
Duplicate
Region
detection

Recombination
block
- Molecular
breeding

# SNV density profile




Homo SNV
Hetero SNV
DP/HK similarity
DP
HK
G.soja
21.0Mb
Gm19 coordinate
1e7

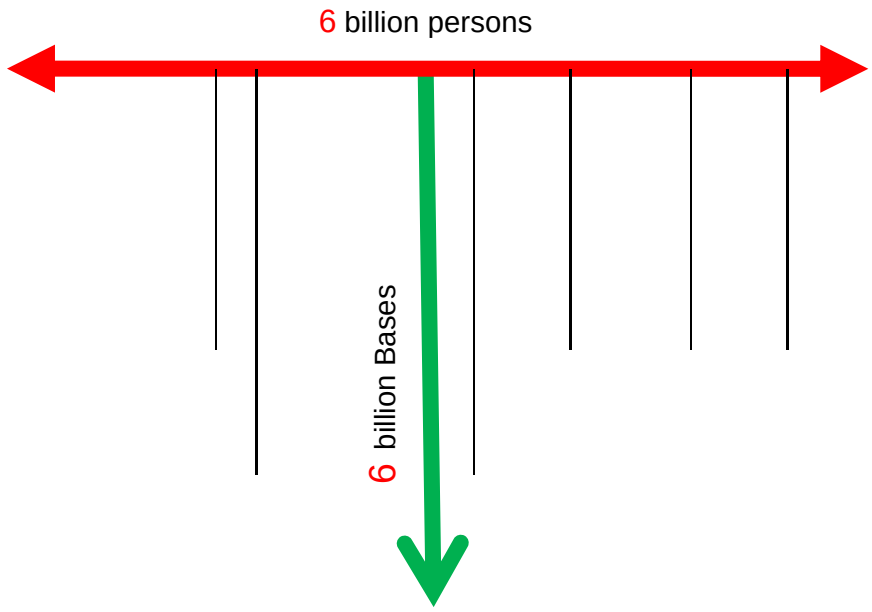

DP/HK similarity
Homo SNV
DP
HK
G.soja
13.8Mb
Hetero SNV
0.7Mb
4.6Mb
Gm20 coordinate
1e7

# 2. Bioinformatics Challenge :

## still massively mapping

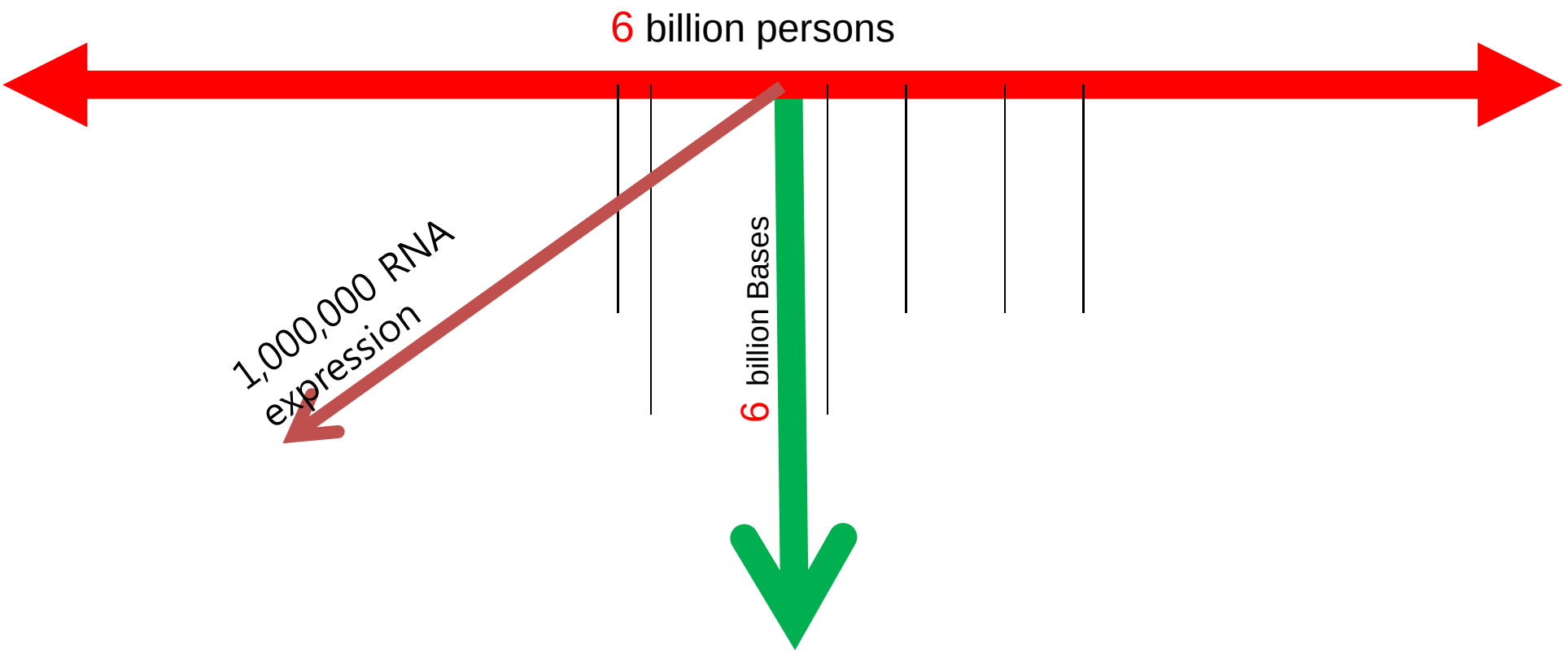
How to map/compute 6 billion X 6 billion matrix?

FullGenome Sequening ➔ Depth Axis
GenoTyping ➔ Width Axis



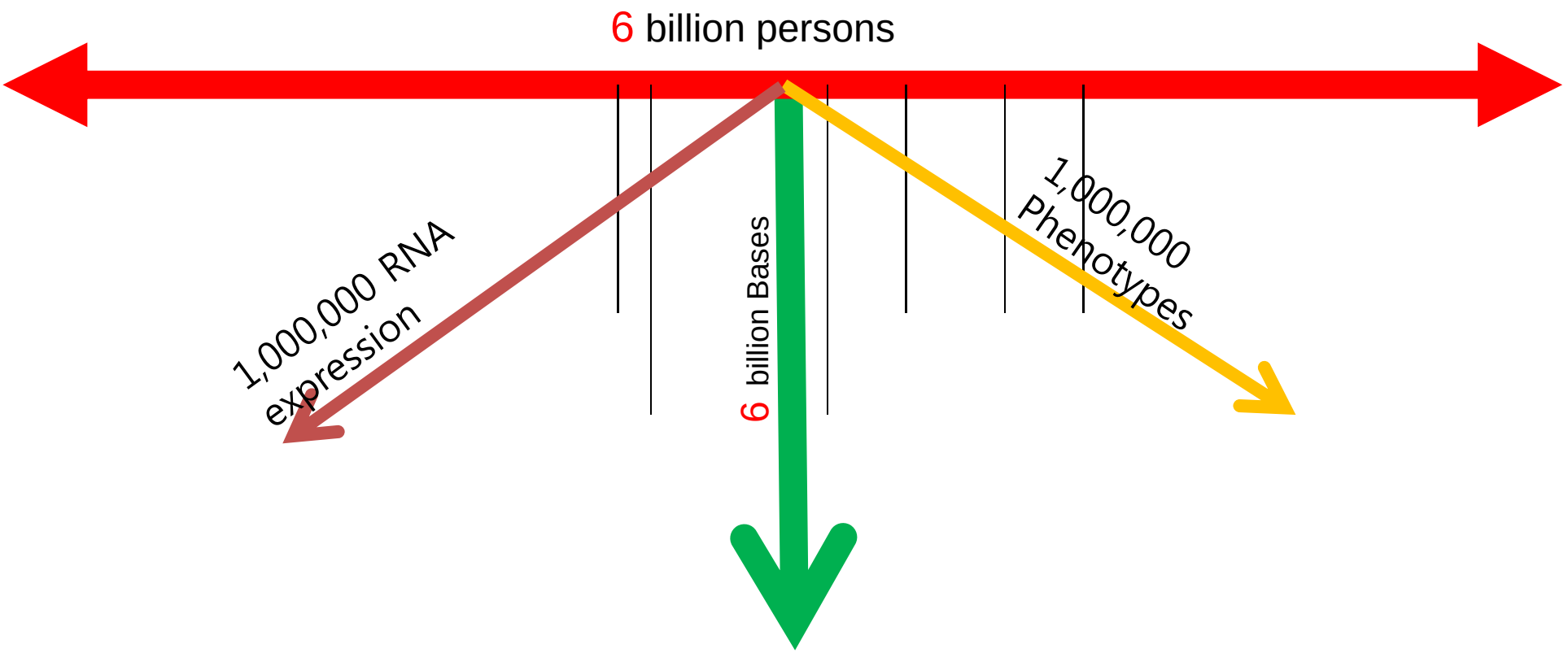
# Adding one more dimension?

How to map/compute **RNA** expressions
In relation with bio-function?

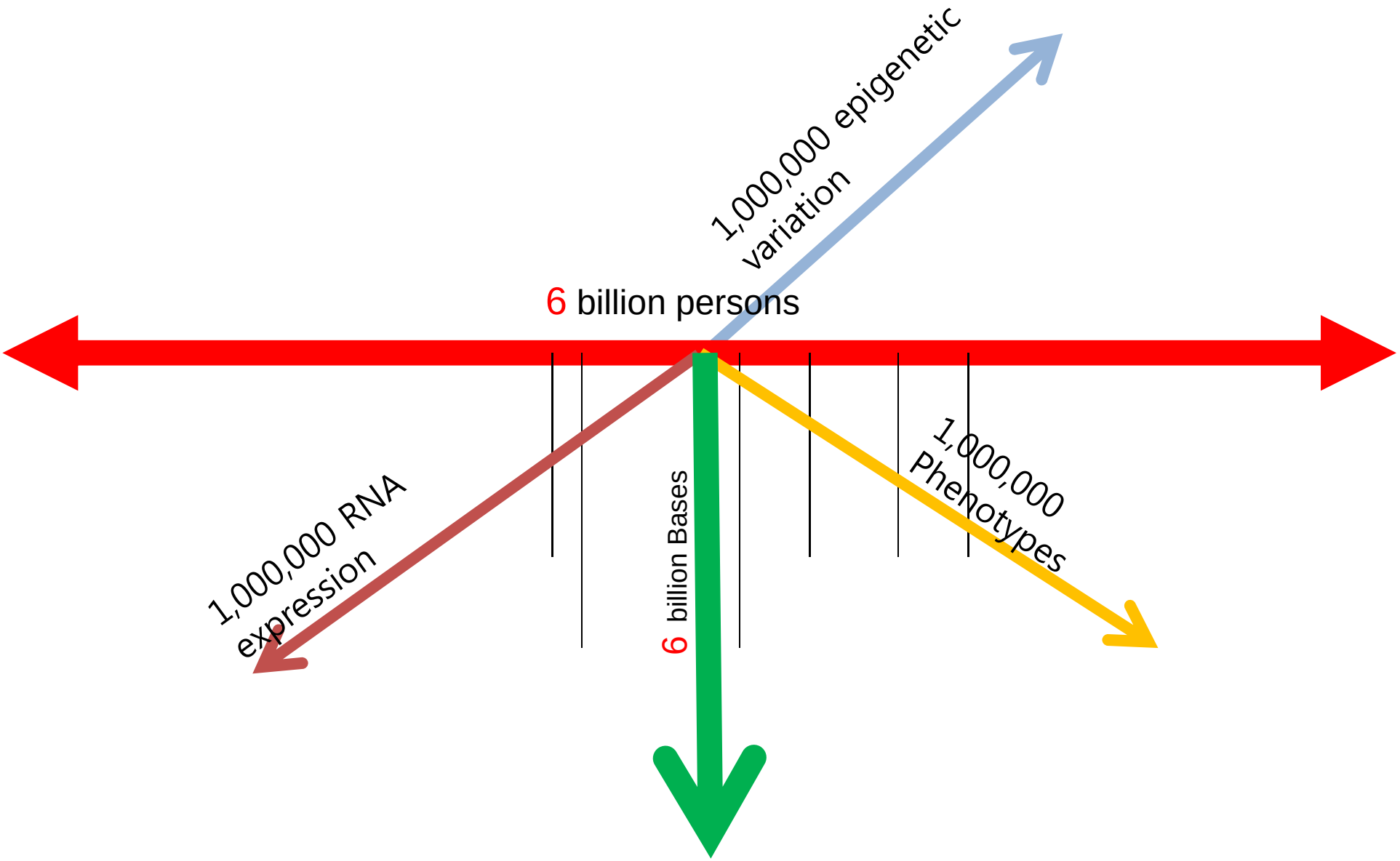


# Adding even more dimension?

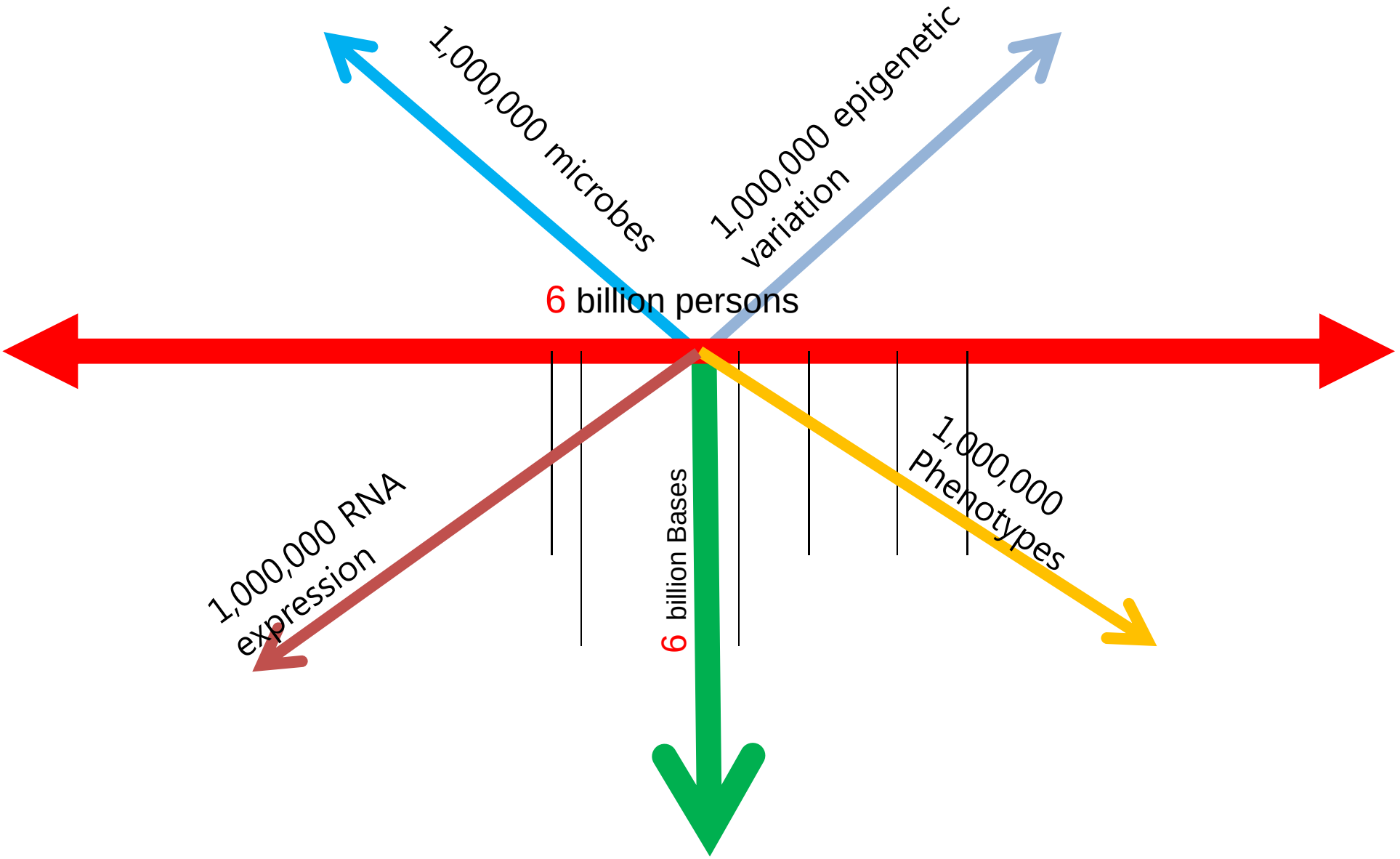
How to map/compute **Phenome?**



# How to map/compute **epigenome**?



# How to map/compute **Microbiome**?



# PGP and KPGP (게놈연구재단 프로젝트)

- "한국인게놈프로젝트"
- Doing Biology with **Sequencing**

# Personal Genome Project (PGP)

- 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

**Mission**

Personal Genome Project is to encourage the development of personal genomics technology and practices that:

- are effective, informative, and responsible
- yield identifiable and improvable benefits at manageable levels of risk
- are broadly available for the good of the general public



# KOREAN PERSONAL GENOME PROJECT

PERSONAL GENOMICS INSTITUTE

THERAGEN ETEX


# GENOME PROJECT 2.0

KOREAN Personal Genome Project

한국인 유전체 지도, 여러분의 참여로 만들어집니다.

REPORT

## 참여현황 »

KPGP의 성장 과정을 보여드립니다.

JOIN

## 참여하기 » 누구나 KPGP와 함께 하실 수 있습니다.

KT선정자분은 실비부담금(100만원)이 면제 됩니다.

유전정보참여 | 연구 참여 | 기업참여

NOTICS

| | |
|---|---|
| 유전정보참여 과정 중, 수집된 임... | 2011-01-04 |
| 유전정보참여는 어떻게 하나요? | 2011-01-04 |
| KPGP에 참여방법은 무엇이 있나요? | 2011-01-04 |
| KPGP참여가 제게 개인적으로 어떤 ... | 2011-01-04 |
| 정부가 아닌 민간에서 KPGP를 수행... | 2011-01-04 |
| KPGP는 무엇을 하는 건가요? | 2010-12-07 |

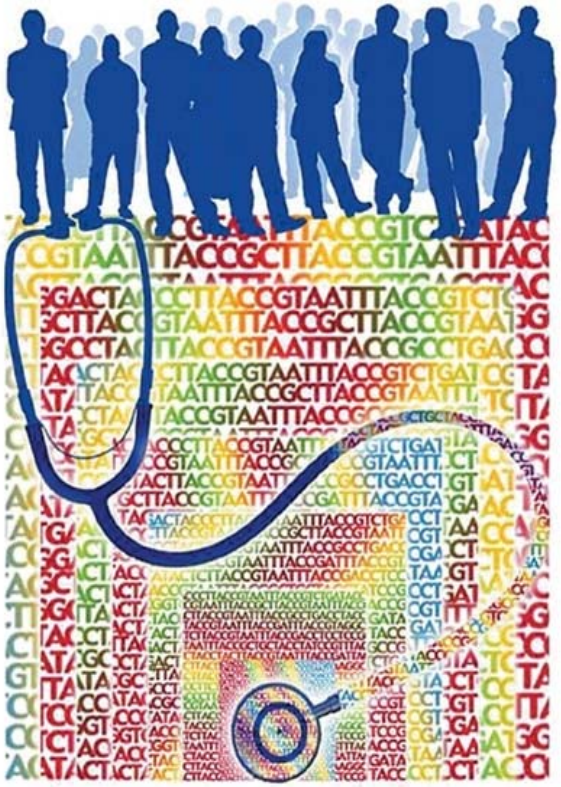

| KPGP 소개 | 연구분야 | 참여안내 | 유전체 이야기 | Community |
|---|---|---|---|---|
| 소개와 비전 | 한국인 표준유전체프로젝트 | 참여안내 | ISSUE&NEWS | 자료실 |
| 미션과 목표 | 희귀질환 표준 유전체 프로젝트 | 유전정보참여 | 참여자이야기 | QnA |
| 보도자료 | HLA형별 참조표준데이터 생산프로젝트 | 기업참여 | | 1:1문의 |
| 연락처 | | 연구자참여 | | |

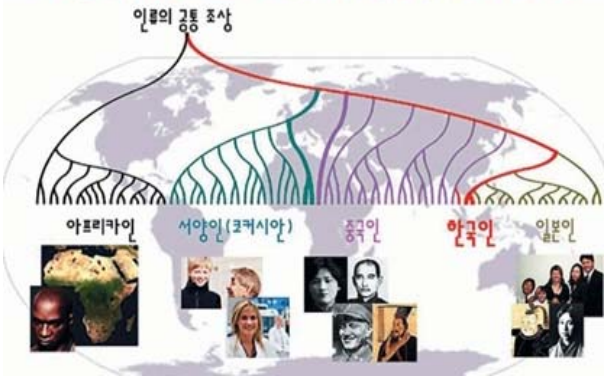



TEL:031-888-9317 FAX:031-888-9314 E-MAIL:info@kpgp.kr

# Korean Personal Genome Project (KPGP)

- Extension of Harvard PGP Project in Korea
- Led by Jong Bhak in Personal Genomics Institute in Korea
- Period : 2007 -2017
- Plan
  - 1단계 2007년 ~ 2009년, 1명 (SJK): Korean Reference
  - 2단계 2010년 ~ 2011년, 100명
  - 3단계 2012년 ~ 2013년, 3000명
  - 4단계 2014년 ~ 2017년, 10000명





# Korean Genome Project Vision & Mission

**Sequencing every single Korean on Earth**

Personalized Medicine using Genomics

**Mission**

1. Korean standard genome information DB
2. Open and sharing genome project
3. New technology development for personal genomics and medicine.
4. Community formation for genomic issues in society (legal, ethical)

# KT 한국인 게놈 프로젝트 참여

olleh

당신이 원하는 그 곳
언제 어디서나 olleh

KPGP
KOREAN PERSONAL GENOME PROJECT

PGI
Personal Genomics Institute

olleh kt

KT, '한국인 게놈 프로젝트' 참여

EVENT 2010/12/06 15:20

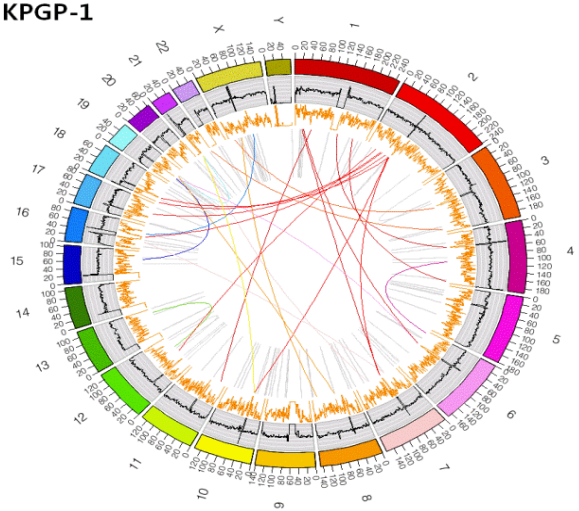

경제
경제일반

KT, 한국인 게놈 연구에 클라우드 컴퓨팅 제공

입력 : 2010-12-06 06:40 / 수정 : 2010-12-06 06:40

# KPGP-20
# Korea Telecommunication (KT) & Theragen


KPGP-1



PGI/TBI
Genome
Technology

KT
Cloud
Technology

20 Normal Korean

Korean Reference Genome

Korean Variation Database

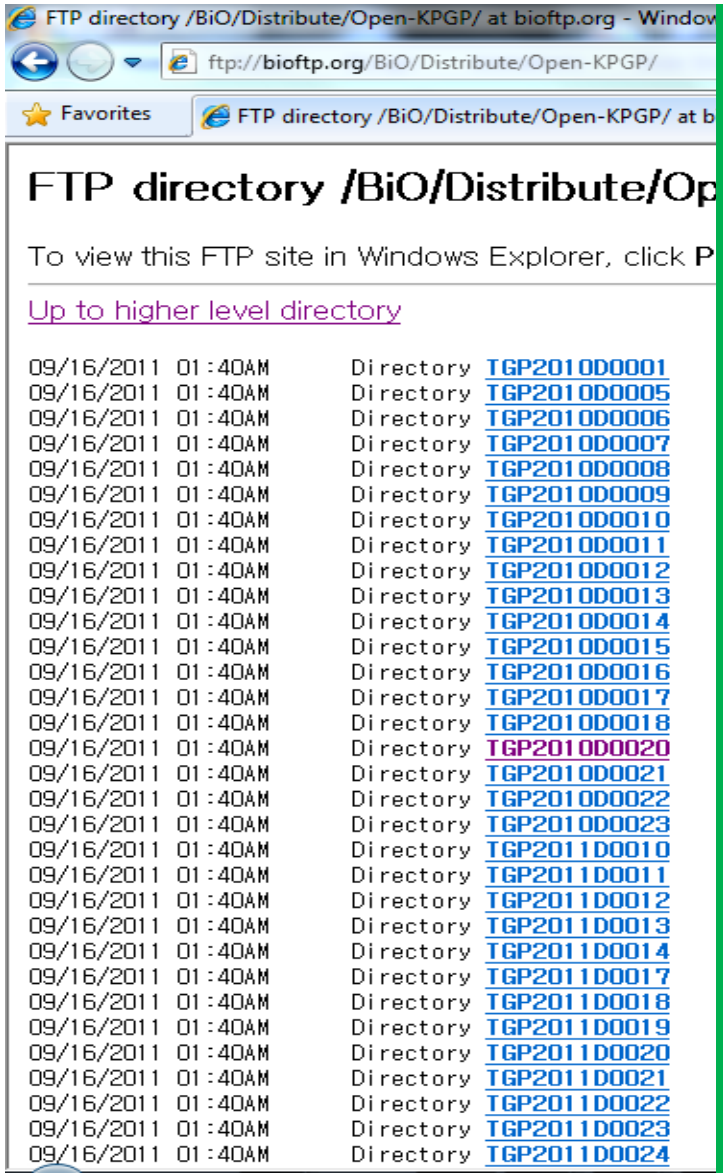
olleh kt

게놈연구재단

PGI
Personal Genomics Institute

THERAGEN ETEX

# Sequencing & Analysis Summary

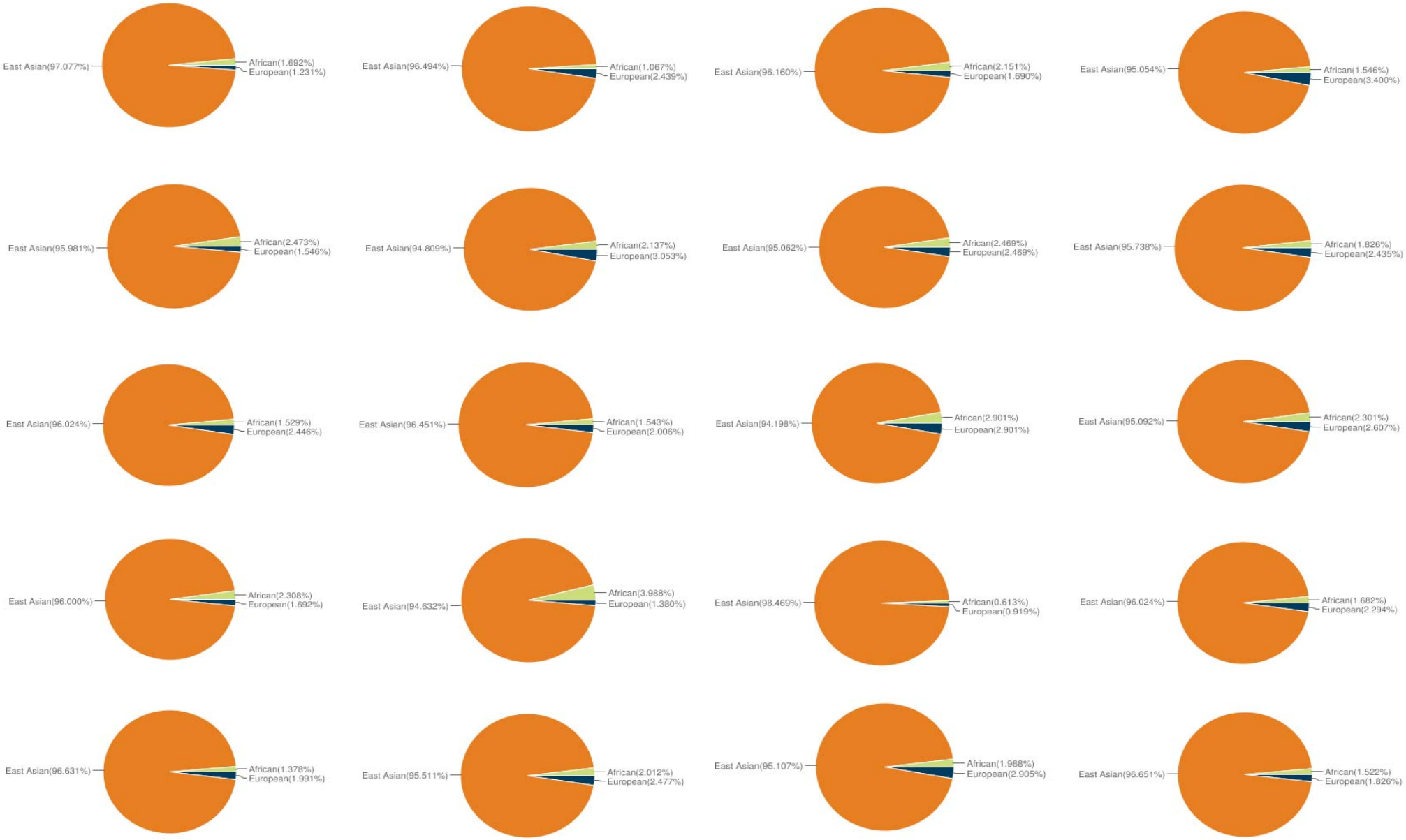



- Sample : Blood Genomic DNA
- Sequencing Platform : Illumina HiSeq 2000

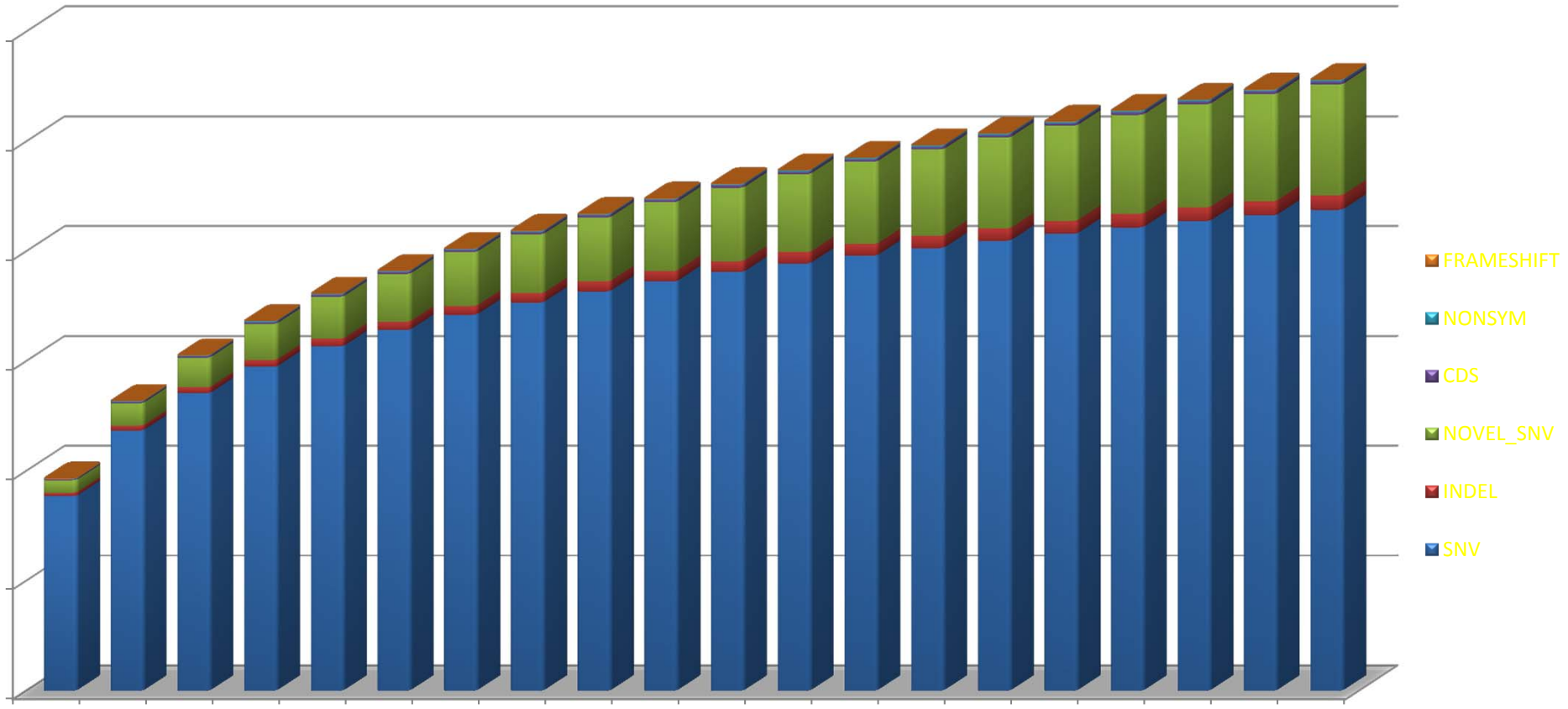
- **Data analysis tool information**
- Raw data fastq
- mapping result bwa(0.5.9)
- SNV samtools(0.1.16)
- INDEL samtools(0.1.16)
- SV breakdancer (1.1)
- NSSNV GRF's pipelin
- STATISTICS GRF's pipeline
- Reference : UCSC HG19

- Open Database :
  **ftp://bioftp.org/BiO/Distribute/Open-KPGP/**

# KPGP-20 Admixture view



# KPGP-20 Cumulative Variation Analysis

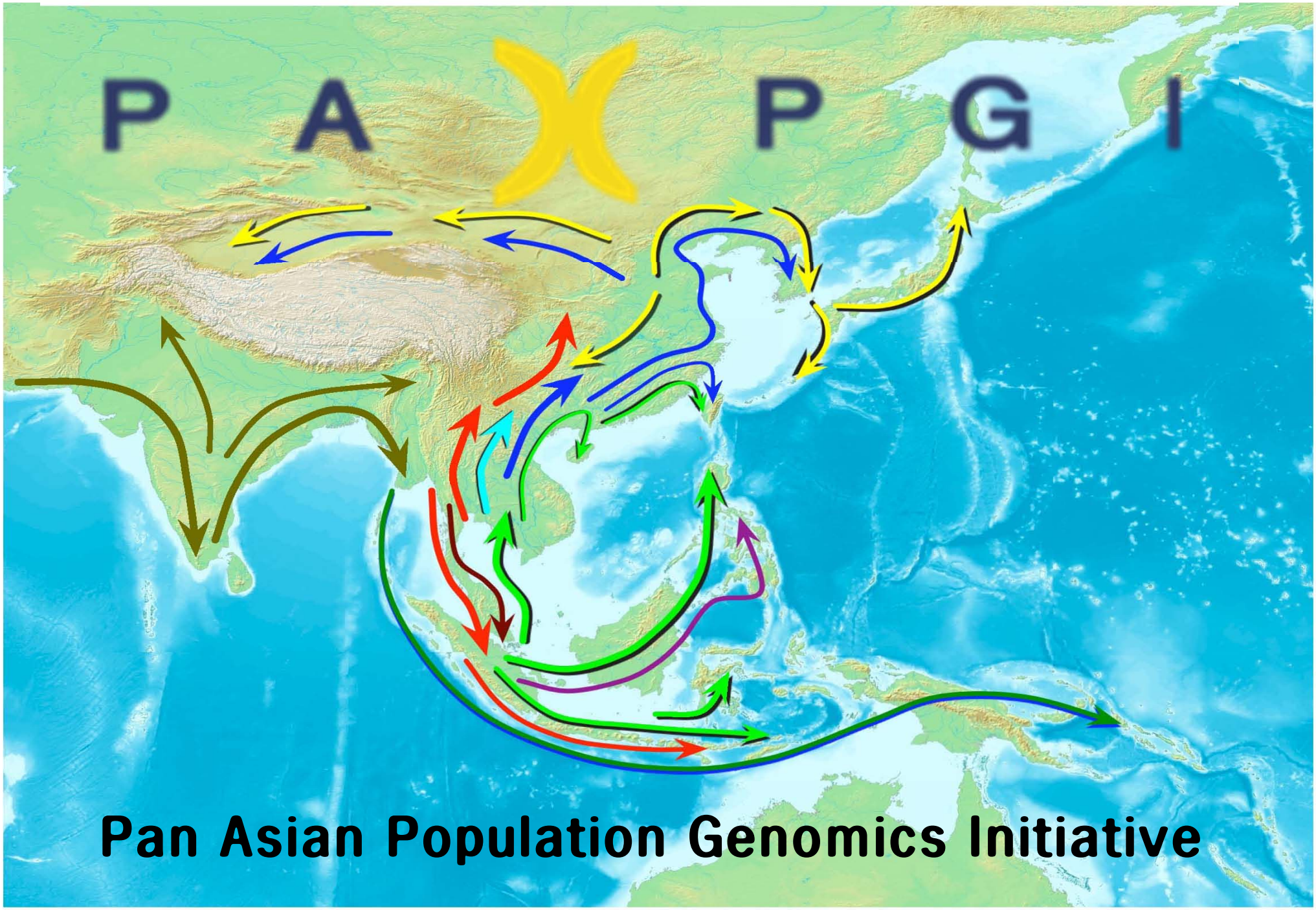


# Sample Size Requirements for Detection

| Number of People (N) | Population Frequency (*f*) | Probability of Detection (p) |
|---|---|---|
| 20 | 0.1 | 0.9852 |
| | 0.05 | 0.8715 |
| 50 | 0.05 | 0.9941 |
| | 0.02 | 0.8674 |
| 100 | 0.02 | 0.9824 |
| | 0.01 | 0.8661 |
| 200 | 0.01 | 0.9821 |
| | 0.005 | 0.8653 |

# PAPGI (PASNP 2.0):
# 범아시아 집단 유전체 프로젝트

# 게놈연구재단


P A X P G I
Pan Asian Population Genomics Initiative

# Useful links

- http://pgi.re.kr
- http://www4a.biotec.or.th/PASNP
- http://papgi.org

- http://opengenome.net

- http://genomics.org
- http://personalgenome.org
- http://personalgenome.net