# Variant Calling from NGS Data

박종화 (Jong Bhak)
(TBI)
테라젠 바이오 연구소
**게놈연구재단**
Jongbhak.com

테라젠이텍스
바이오연구소

# BioAcknowledgement


- Researchers who are **honest** and **passionate** in doing science
- People who support scientific research by **paying taxes**
- TBI & Genome Research Foundation colleagues
- PGI and TBI friends.
- SKT

# 결론 1.

# 해독을 해야합니다.

# 결론 2.

## 게놈을 많이 해독해야합니다.

## 한국인게놈프로젝트:
## 5,000 만명 게놈 프로젝트

# 결론 3

## 게놈을 다 해독 해야합니다.

## “70억명 게놈 프로젝트”

**http://billiongenome.com**

# 결론 4:

# “게놈 기본권리”

## 만듭시다

http://checkgenome.co

# 결론 5: 게놈법

# "게놈산업 육성법"

Genome research boosting law

# 게놈 이야기

## 게놈은 “궁극산업”의 하나

* 궁극산업: 인간의 몸으로는 궁극적으로 필요한 산업

**세상의 모든 것은 게놈으로 연결되어 있다**
**세상의 모든 공간은 게놈으로 채워져 있다**
**세상의 모든 일은 게놈에 의해 처리된다**
**세상의 모든 산업은 게놈정보를 쓰게 된다**

*** 정보산업 VS 게놈산업**

# 게놈 혁명합시다!

비영리 **게놈연구재단**

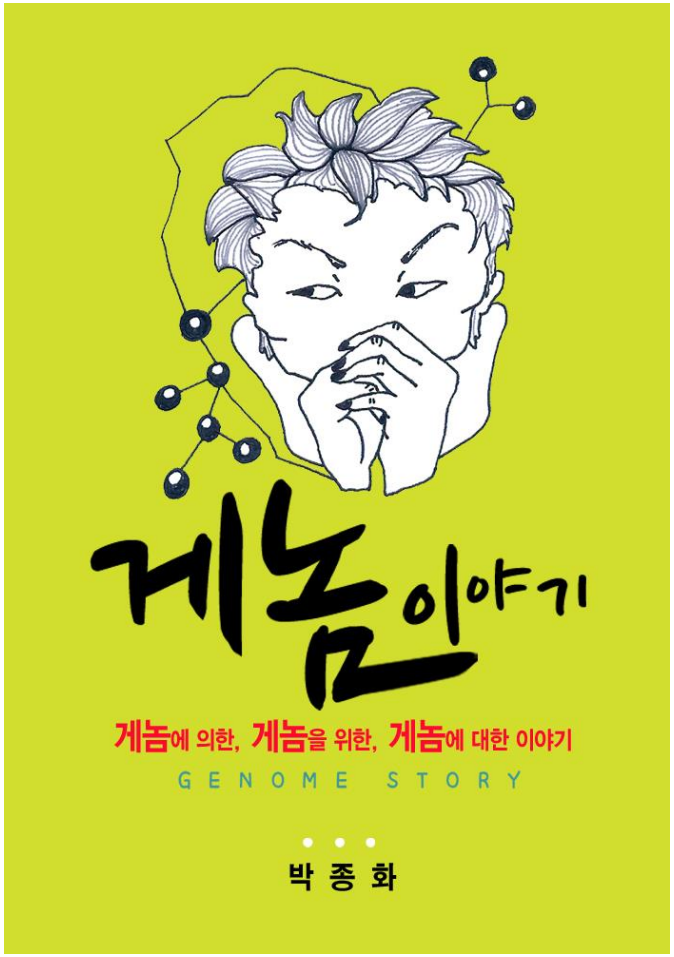
By Jong Bhak

# 게놈연구재단
# (Genome Research Foundation)

- 비영리 연구재단 (2010년 4월 19일 게놈연구에 뜻이 있는 사람들이 테라젠과 고진업대표의 지원으로 설립)
- 목적: 죽을때까지 게놈을 연구하고, 사랑하고, 정리하고, 팔고 사고 하는 곳
- 기부/연구용역 받습니다 (돈좀 주세요)
- 연구진:
  - 최초 한국인 게놈 분석
  - 최초 호랑이, 사자, 눈표범 게놈 해독
  - 백호랑이 색깔(2013년 Current Biology)
  - 최초 고래게놈(해양연 공동 프로젝트)
  - 위암, 폐암연구

# 게놈이야기
(자서전)

- **http://www.upaper.net/**
- http://www.upaper.net/biomatics/1009617
- 저자:
  변하나, 김진섭, 박종화



# 목 차

## Variant Calling

0. 게놈/생정보학 서론 (intro for Genome and Bioinformatics)
1. NGS workflow (차세대 해독 워크플로)
2. DATA Format (데이타 포맷)
3. 암 샘플에서의 Variant calling의 차이점

**The universe?**

(우주란 무엇인가)

**It is a sequence**

(서열이다)

# Sequence?

(서열이란?)

# What is genomics?
# (게놈학이란?)

참고: http://genomics.org

http://omics.org

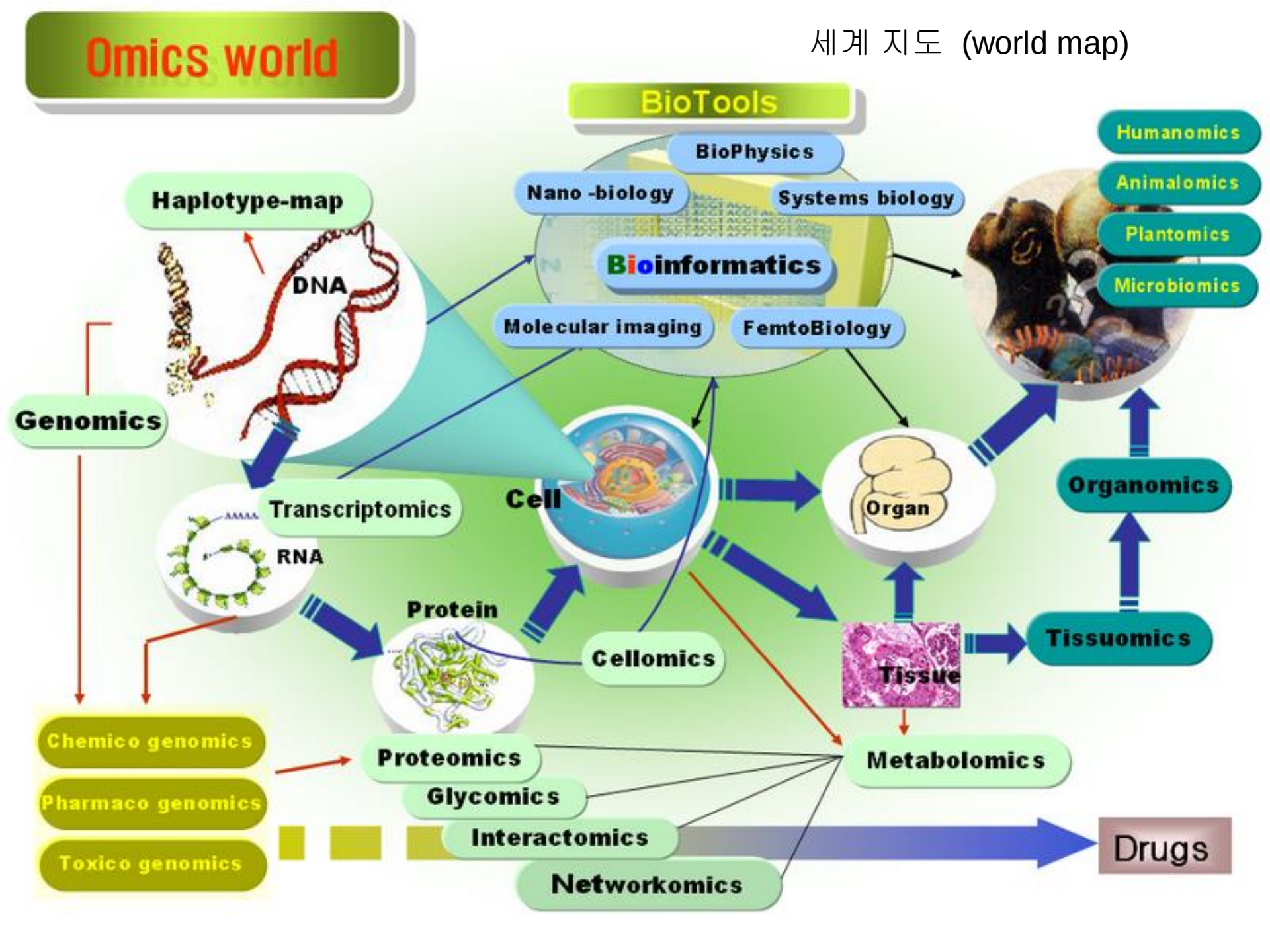
# 세계 지도
# (world map)

강니도
1402
(태종)
좌정승 김사형, 우정승 이무와

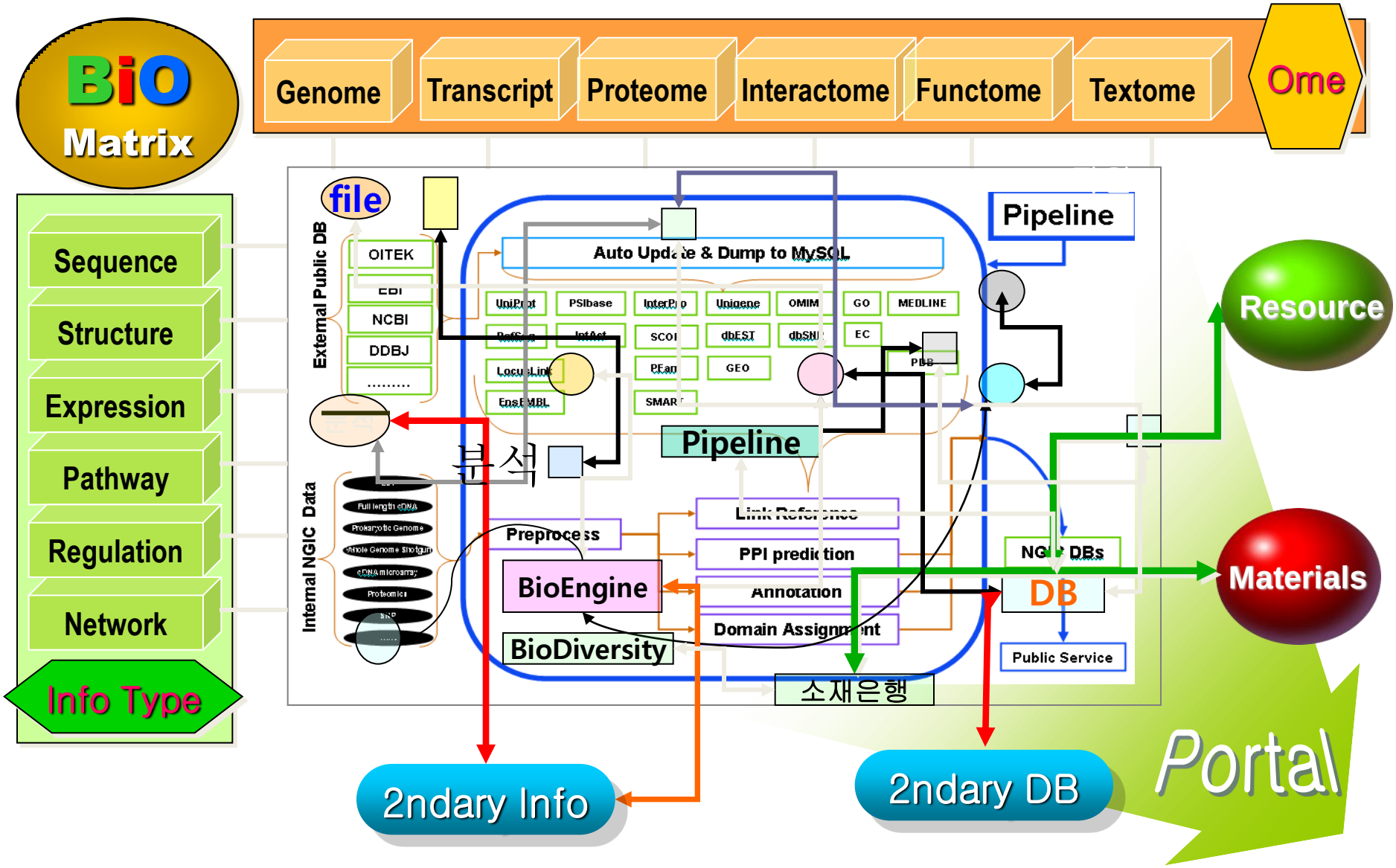
이회가 만든 세계지도

# Omics world

세계 지도 (world map)



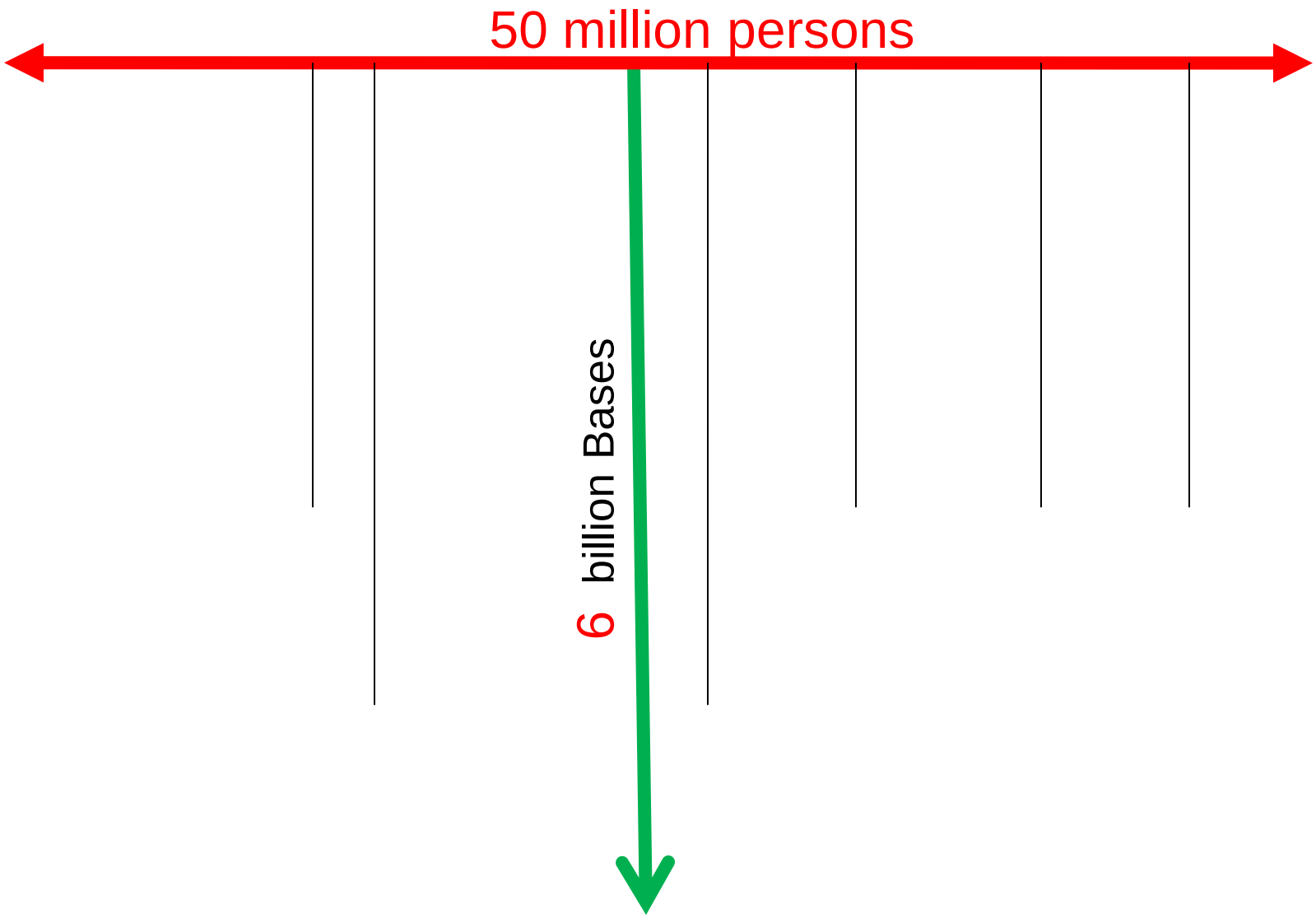
# BioMatrix: Putting structure in omics



# 게놈의 포인트?

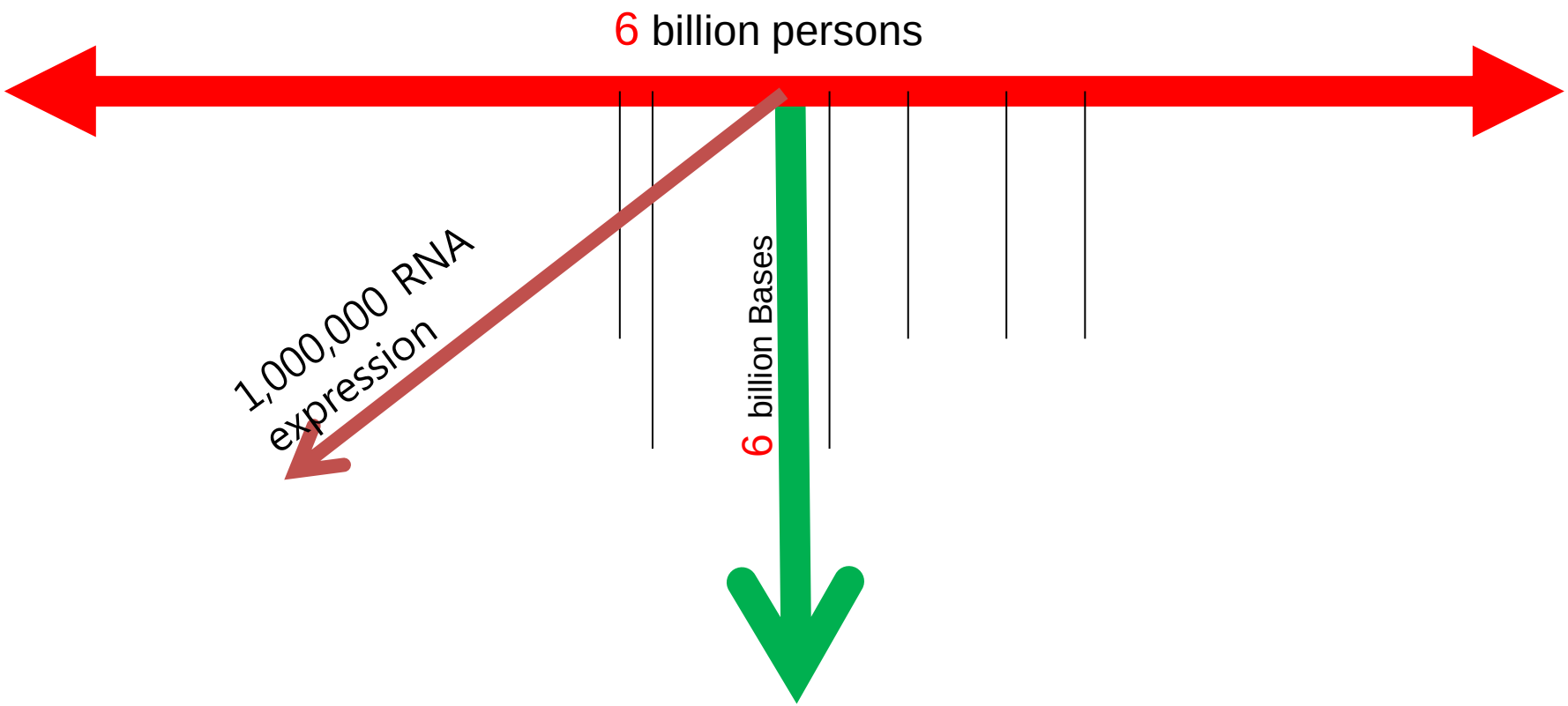
- 변이
- 변이체 (Variome)

# 게놈 T :



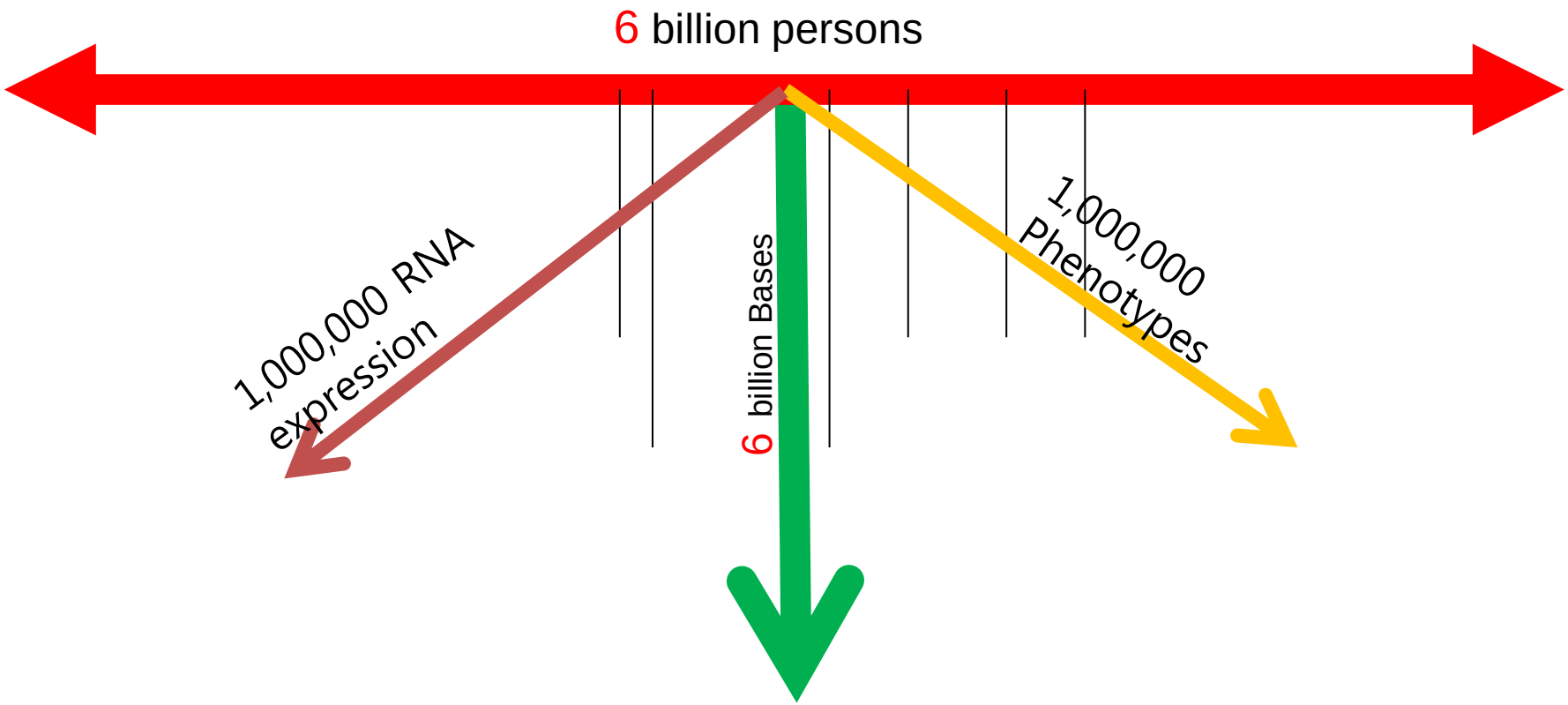
# Adding one more dimension?

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

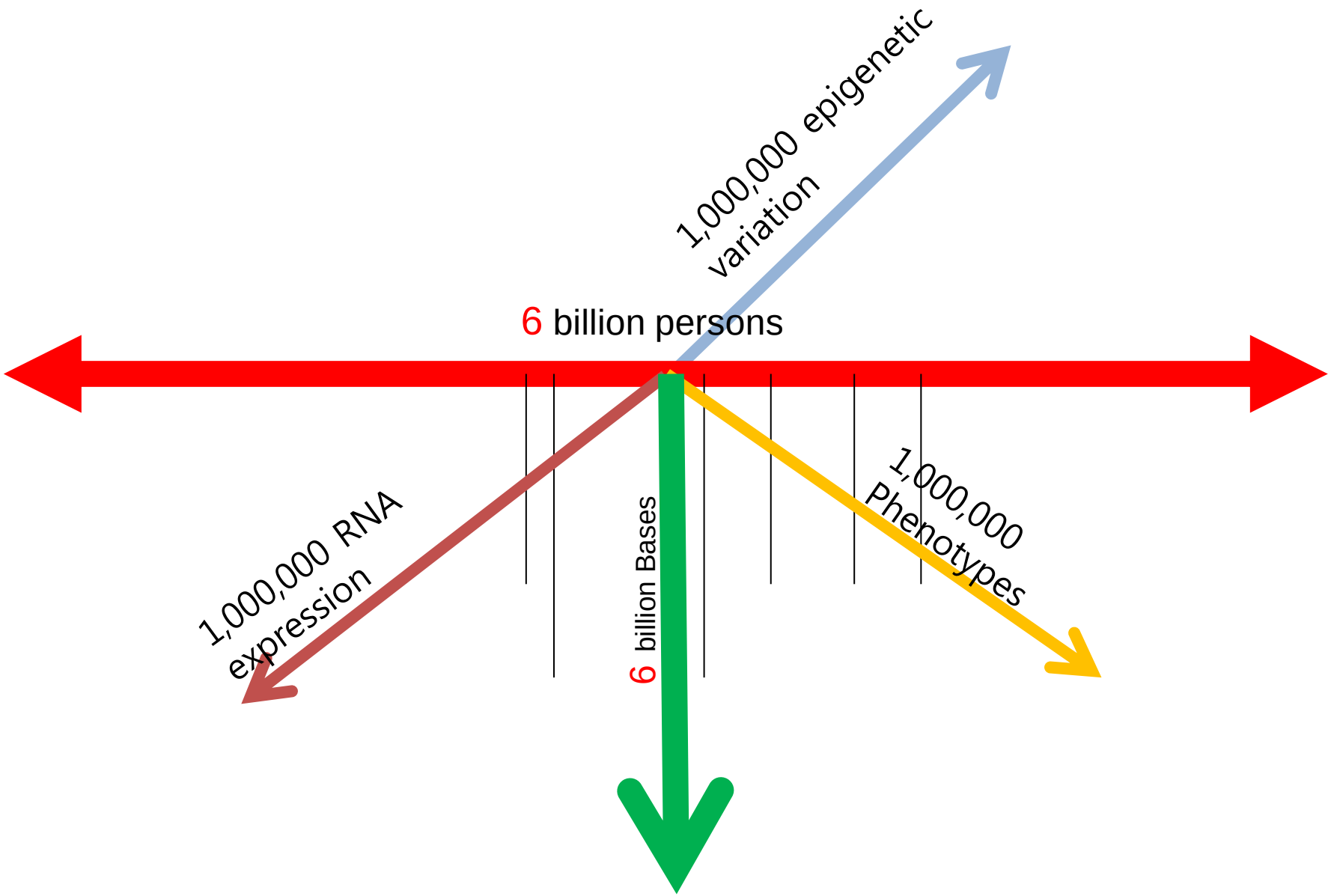


# Adding even more dimension?

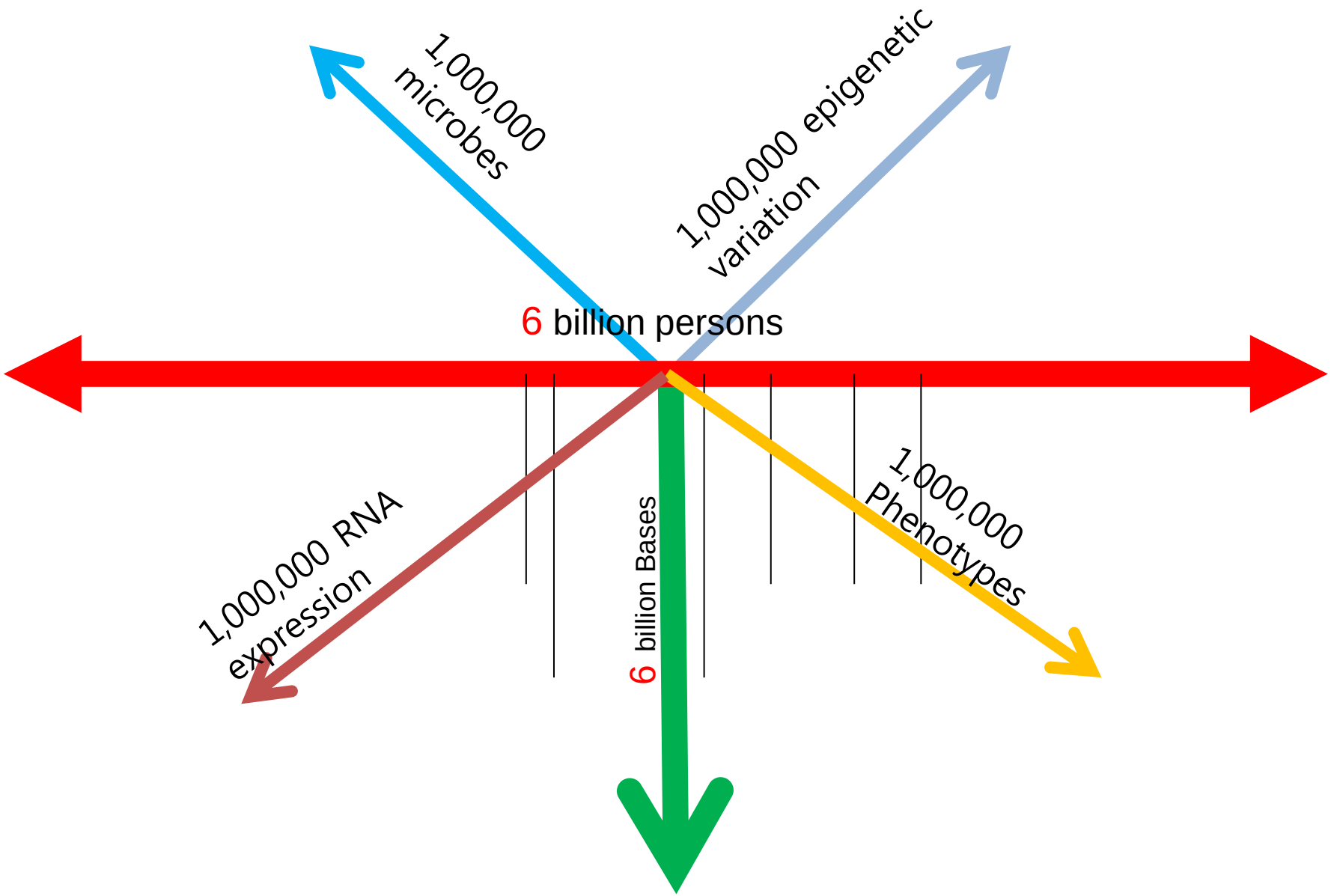
How to map/compute **Phenome?**



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



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



# Variant Calling Process


- ✓ **Aim**: produce variant calls and genotype calls

- ✓ **Difference between variant site and genotype**:
  - ref is A, aligned bases are TTTTTTTAA
  - highly likely that the site is variant
  - less clear what the genotype is: T/A or T/T?

# Variant Calling Issues

- ✓ Primary data
  - PCR errors (base errors)
  - Base quality calibration
  - Indel errors
  - Overlaps
  - Duplicates
  - Primers
- ✓ Alignment
  - Base-level mis-alignments around Indels
- ✓ Reference / mapping
  - Unrepresented segdups
  - Repetitive sequence

# Cancer

**암 샘플에서 VARIANT CALLING 의 차이점**

# Mutation 종류

- **Germline** mutation
  - Hereditary
- **Somatic** mutation
  - Acquired

# Germline mutation

- Germ line cell 에 발생된 변이
- 신체의 모든 세포가 이 변이를 공유함
- 다음 세대로 유전됨

- 확인방법: 부모에게 없는 변이
  - 개인 게놈 분석시 발견되는 ~400만개의 SNV 대부분은 그 개인의 조상의 germline mutation의 축적에 의한 것임
- 활용: 가족력이 있는 유전질환 연구의 대상

# Somatic mutation ?

- Germ cell 이외에 세포에 발생하는 변이
  - 모든 세포의 염색체가 모두 다를 수 있음
- 유전되지 않음 (?)

- 발생시기에 따라 분포하는 범위가 다름
- 발생원인에 따라서 개수가 다양함
  - 자연적 발생: 수십~수백 (모든 세포에서 발생하는 모든 somatic mutation의 개수는 DNA polymerase의 error rate로 계산 가능)
  - 암: 수만~수십만개 (분리된 암조직에서 추출할 경우)
- 활용: 암 등의 원인변이 연구

# Somatic mutation

- **Spontaneous** : 정상적 세포활동에 의해 발생하며, 각 단백질이 100% 완벽하게 작동하지 않아서 생기는 문제
  - DNA replication error
  - DNA repair error
    - Mismatch repair
- **Induced by mutagens**
  - UV
  - Chemicals
  - Virus
  - …

# SNV 구성

- 1인당 약 400만개의 SNV 찾아짐 =

  **조상의 germline mutation**에 의해 축적된 변이중 일부
  (400만개 거의 대부분)

  ＋ (더하기) **자신의 Germline mutation**
  (수십개?)

  ＋ (더하기) **자신의 somatic mutation** 중 일부
  (소량)
  (예. 샘플이 혈액일 경우 백혈구를 만드는 조혈모세포의 변이)
  (예. 샘플이 암조직이라면, 암조직으로 분화되며 생긴 변이)

# NGS에 의한 암 조직의 Somatic mutation

암 조직의 Somatic mutation =

total SNV 400만개 – (빼기) 정상조직 변이 (germline mutation + 정상조직의 somatic mutation)

예) 고형 암조직 somatic mutation 추출

암조직 변이 – (빼기) 정상 인접조직 변이

암조직 변이 – (빼기) 혈액 변이

예) 백혈병 somatic mutation 추출

혈액 변이 – 정상조직 변이 (구강상피세포 등)

# **암 조직에서** Somatic mutation calling의 어려운 점

- 해독 및 분석의 정확성
  - SNV calling 자체의 정확도 한계

- 샘플의 purity
  - Stromal admixture
    - 외과적 분리시 암조직만 100% 분리하기 어려움
    - 순도가 아주 낮을 경우 SNV calling 시 noise와 분리하기 어려움
  - Clonal evolution
    - 암세포의 mutation 축적에 따른 분화
    - 한 개의 암조직에 다양한 종류의 암세포가 섞여있음

# 암 분석 문제의 해결

- 샘플
  - Microdissection 등을 통해 purity 확보
  - High depth sequencing

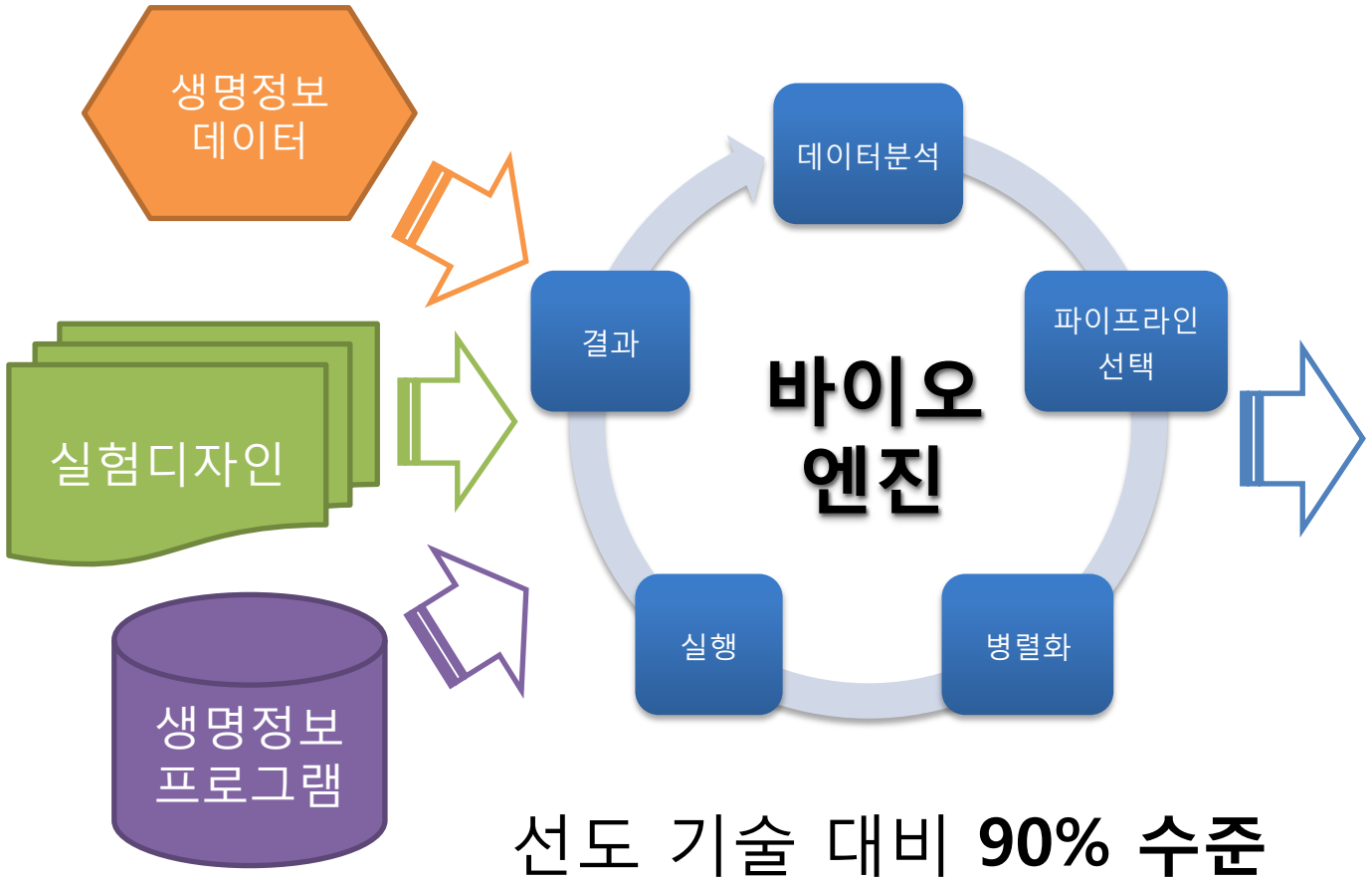
- 해독 및 분석
  - 정확한 mapping, SNV calling 알고리즘
  - 고 효율 false positive filtering 알고리즘

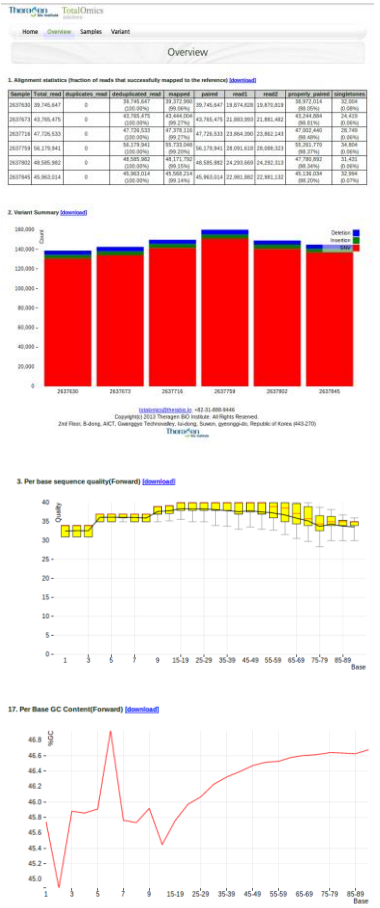
(테라젠은 10% purity에서도 90% 정확도로 분리 가능)

# 게놈 산업?

# 제안: 산자부과제: GiSys 시스템

데이터 → 정보분석 → 레포팅




선도 기술 대비 **90% 수준**

# 이쓔

## 누가 가장 많이 “개인 게놈”으로 돈을 벌것인가?



















# 한국에서?

테라젠 ?
DNAlink ?
마크로젠 ?
삼성 SDS ?
LG ?
SK ?
KT ?

# 생명공학의 수익창출

- 생명정보사업(소프트웨어)
  - 검색, 병원시스템, 연구정보 활용, 지식데이타베이스
- 생명공학사업(하드웨어)
  - 제약, 건강식품, 농수산물
  - 컴퓨터, 네트워크, 스토리지

# 한국의 게놈분야 상황

- 생산 원천기술: 없음
- 대량생산체제: 진행중
- 정보분석인프라: 거의 없음 ➔ GiSys

# 결론

- BioRevolution 이 진행되고  있음
- **정보기술**이 생명공학 혁명을 주도함
- 한국과 생명공학 혁명:
  - 제2의 반도체산업 혹은 그 이상
- 한국: 미래의 기회: 불확실 ➔ 기회
- 투자 제안
  - **대량 바이오칩/시퀀싱 생산기술에 투자**
  - **개인게놈 파생상품회사에 투자**
  - **좋은 생명의료정보 해석/분석 회사에 투자**