BiG Genomics

(Billion Genome Project)

Jong Bhak

Theragen BiO Institute

San Diego

BioAcknowledgement

- Researchers who are honest and passionate in doing science
- People who support scientific research by paying taxes
- TBI & Genome Research Foundation colleagues
- Maryana Bhak for support and editing
- NSF and Jan/Olga Vitek for organizing this conf.

BioDisclaimer

- Everything I present here is stolen from all other human beings and machines dead, alive, and will be alive.
- Whenever I say ‘I’, it means all of us or We.
- There is no copyrights whatsoever here and everything is under copytheft.
- All the contents here is under **BioLicense** (i.e., it is all yours): http://biolicense.org

Conclusion 1.

Let’s do more sequencing!

Post-genomic? ➔ Genomics era has not arrived yet.

By Jong Bhak

Conclusion 2: Even more

Let’s sequence every Korean as cheaply as possible.

“50 mil. Korean Genomes”

By Jong Bhak

Conclusion 3: Everyone

Sequence 7 billion people on Earth as fast as possible and analyze them.

“7 billion Genomes”

http://billiongenome.com

By Jong Bhak

Genome Law

Genome research stimulus law

“Genomics Bill”

Genome Rights

Everyone has the right to know his/her own genome information

“Genome Bill of Rights”

http://genomerights.org

Big Data?

- Earth is a big network of distributed computers ➔ They are processing some data.

- These computers process a massive amount of biological and environmental data.

Any big data?

- Genomes and derivations are **the only ‘big’ data** we have on Earth. ☺

Terms

- **Big data ➔** Massive amount of genomic data, a neologism for getting grants.
- **Cloud ➔** Big server for analyzing genomic data, a neologism for getting grants.
- **Programming ➔** Communicating with our brains that reside out side of our skulls, a name for something we have been doing for about the past 4 billion years.

Programming?

- Talking to ourselves.

- Best programming language

  ➔ English

Programming with Big Data?

- Talking to ourselves about genome data.

Programming with Big Data?

- Talking to ourselves, using silicon based brains, about Genomic data.

- Talking to ourselves, using silicon based brains in English on the net, about **next generation sequencing** derived Genomic data.

- Talking to as **many of us** as possible, using silicon based brains called **computers** in English on the net, to process next generation sequencing derived **Genomic, Proteomic, and Metabolic** data to understand the **structure of information.**

Programming with Big Data?

- Talking to as many of us as possible, using silicon based brains called computers in English on the net, to process next generation sequencing derived Genomic, Proteomic, and Metabolic data to understand the structure of information that will help us live longer and conquer cancers, diabetes, flu, Alzheimer’s, and asthma.

To do well in PwBD

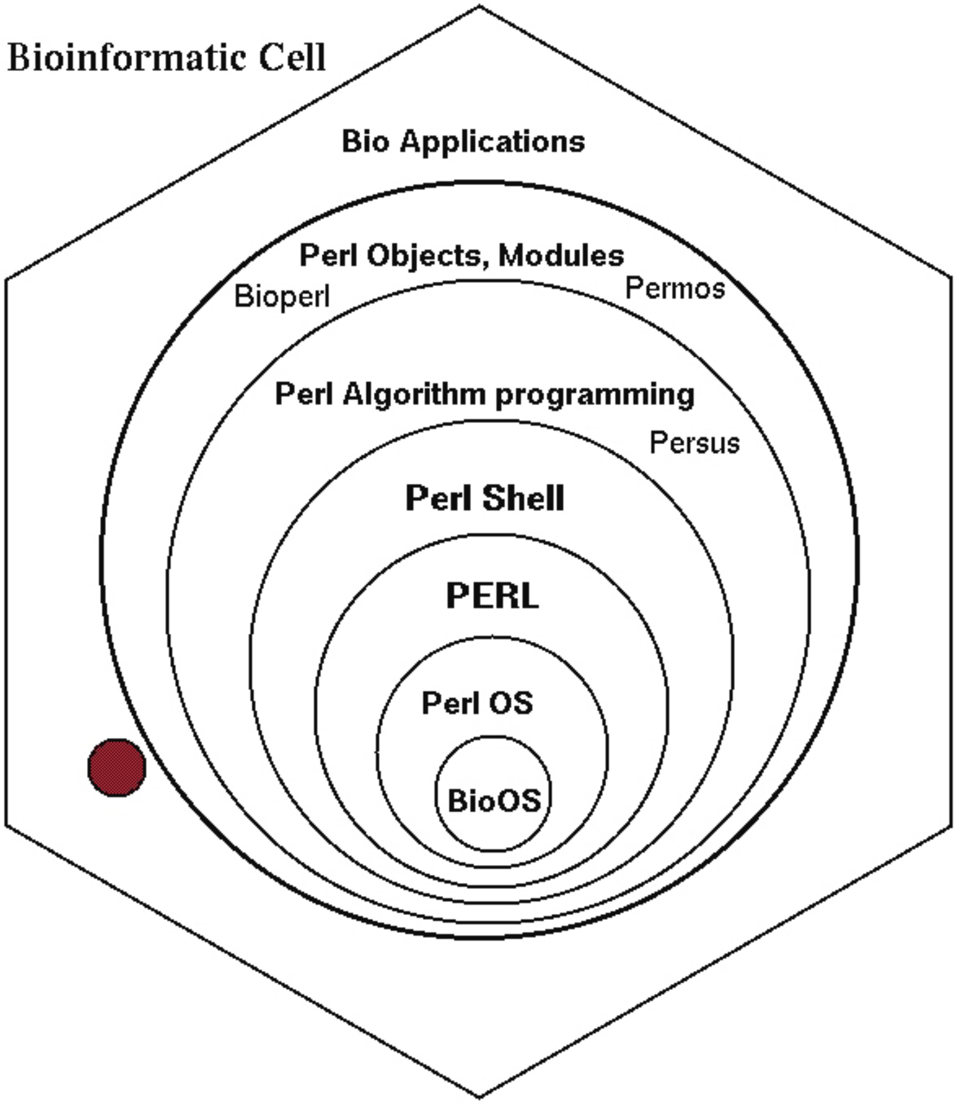
- **Talking to as many of us as possible ➔** come to Hawaii often.
- **using silicon based brains called computers ➔** buy many computers using NSF grants.
- **in English ➔**
- **on the net➔**
- **to process next generation sequencing derived Genomic, Proteomic, and Metabolic data ➔**
- **to understand the structure of information ➔**
- **that will help us liver longer and conquer cancers, diabetes, flu, and asthma**. **➔**

Genome

- Genome is a self-coding language / program

- It is not the Operating System
  - It needs an OS, compiler, middleware, shell, IDE, visualizer, pipelines, and applications

The Bioinformatic Cell: 1999

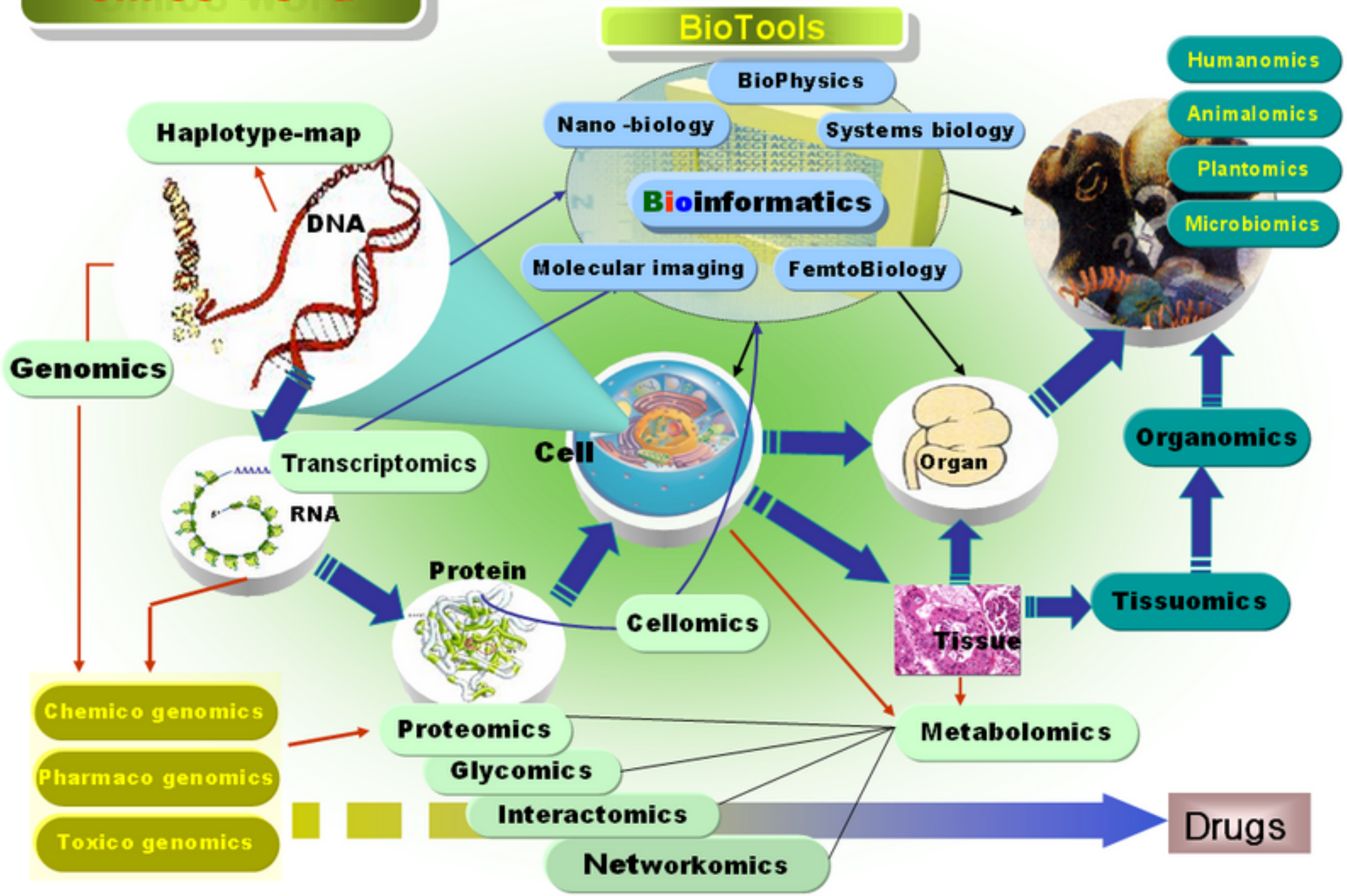


Bio[.+]

- BioOS   BioLinux
- BioPerl/BioJava/BioPython/BioRuby/ BioPHP
- BioProgramming..

http://bioprogramming.org
http://biolinux.net
http://bioperl.net
http://biophp.net
http://bioos.org
http://biojava.net

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

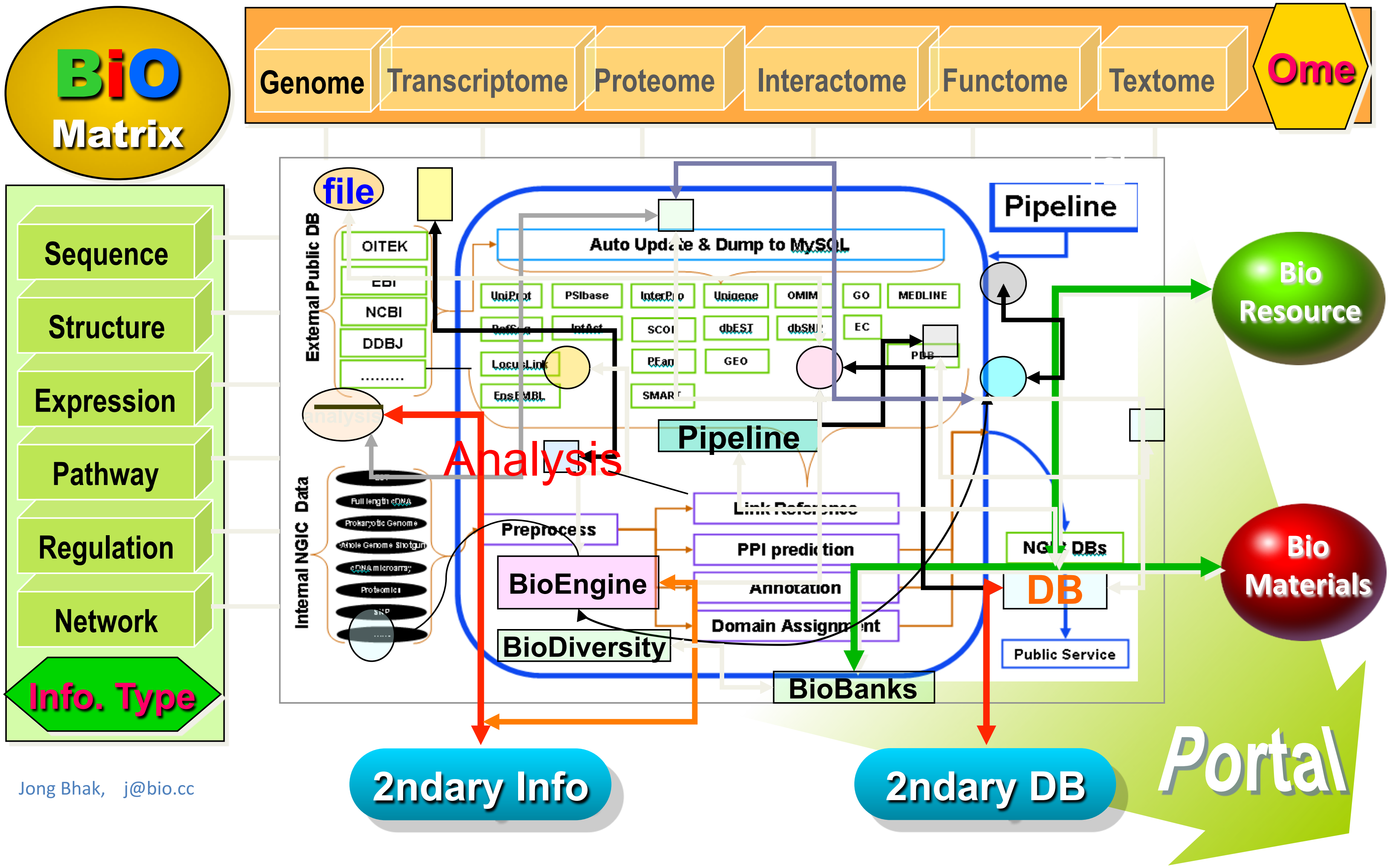
Interactomics

Networkomics

Metabolomics

Drugs

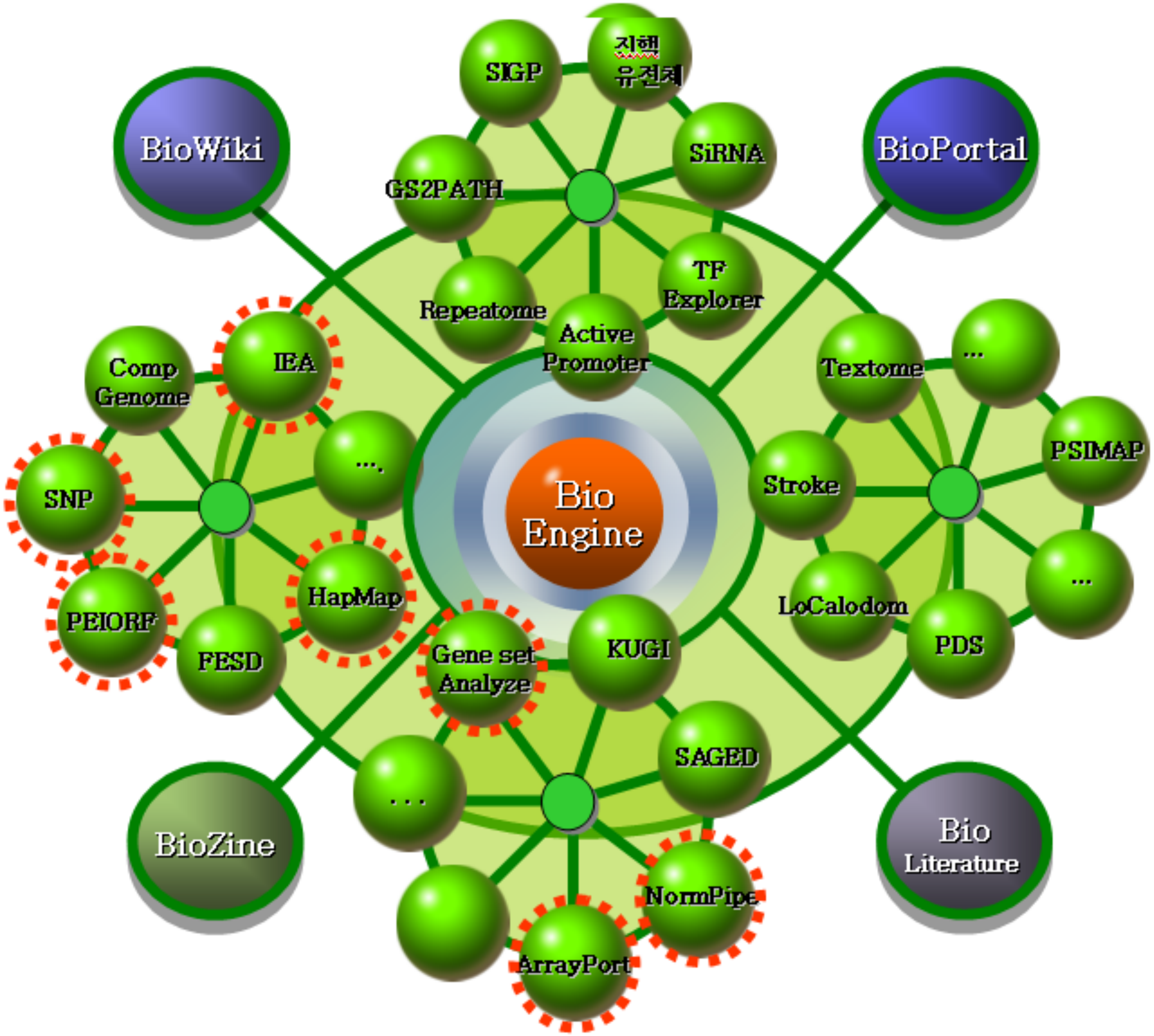
Hacking The Biomatrix



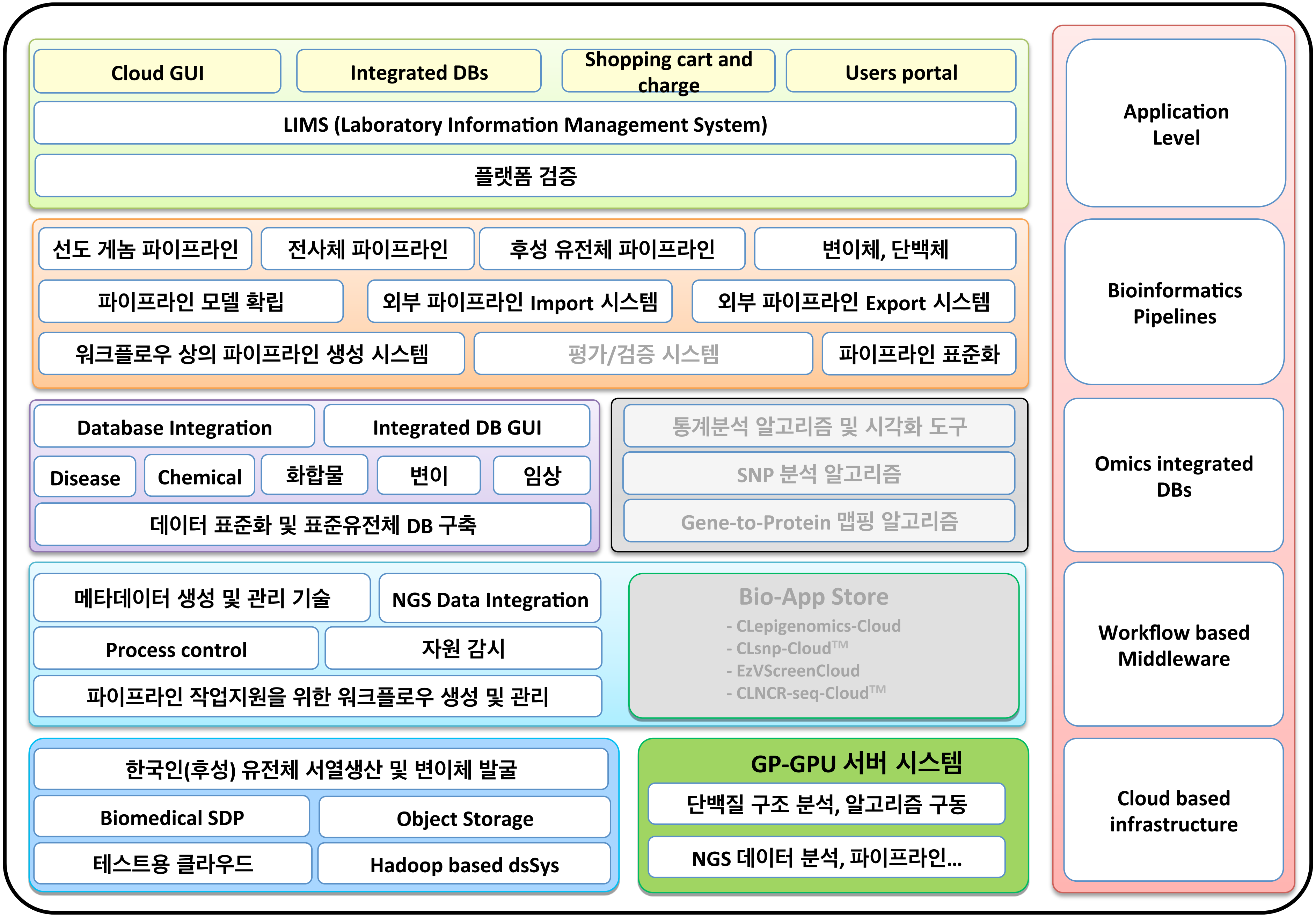
Jong Bhak, j@bio.cc

Jong Bhak, 20051202. j@bio.cc, http://bioinformatics.ws/, under BioLicense

BioEngine: Automatic BioInformation Pipelines Processing System



GiSys



Cloud GUI

Integrated DBs

Shopping cart and charge

Users portal

LIMS (Laboratory Information Management System)

플랫폼 검증

선도 게놈 파이프라인

전사체 파이프라인

후성 유전체 파이프라인

변이체, 단백체

파이프라인 모델 확립

외부 파이프라인 Import 시스템

외부 파이프라인 Export 시스템

워크플로우 상의 파이프라인 생성 시스템

평가/검증 시스템

파이프라인 표준화

Database Integration

Integrated DB GUI

Disease

Chemical

화합물

변이

임상

데이터 표준화 및 표준유전체 DB 구축

통계분석 알고리즘 및 시각화 도구

SNP 분석 알고리즘

Gene-to-Protein 맵핑 알고리즘

메타데이터 생성 및 관리 기술

NGS Data Integration

Process control

자원 감시

파이프라인 작업지원을 위한 워크플로우 생성 및 관리

Bio-App Store

- CLepigenomics-Cloud
- CLsnp-Cloud™
- EzVScreenCloud
- CLNCR-seq-Cloud™

한국인(후성) 유전체 서열생산 및 변이체 발굴

Biomedical SDP

Object Storage

테스트용 클라우드

Hadoop based dsSys

GP-GPU 서버 시스템

단백질 구조 분석, 알고리즘 구동

NGS 데이터 분석, 파이프라인...

Application Level

Bioinformatics Pipelines

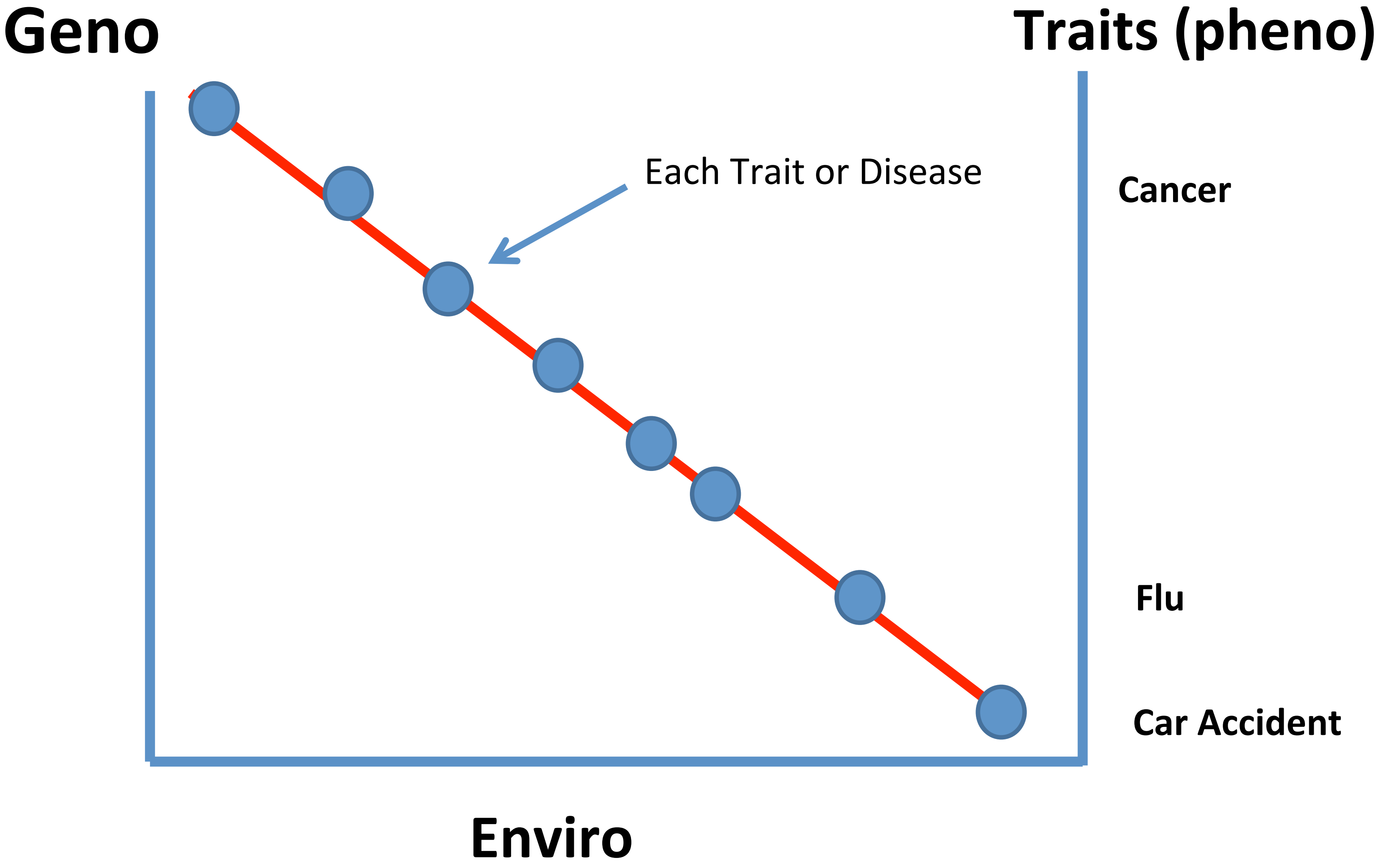
Omics integrated DBs

Workflow based Middleware

Cloud based infrastructure

To do What?

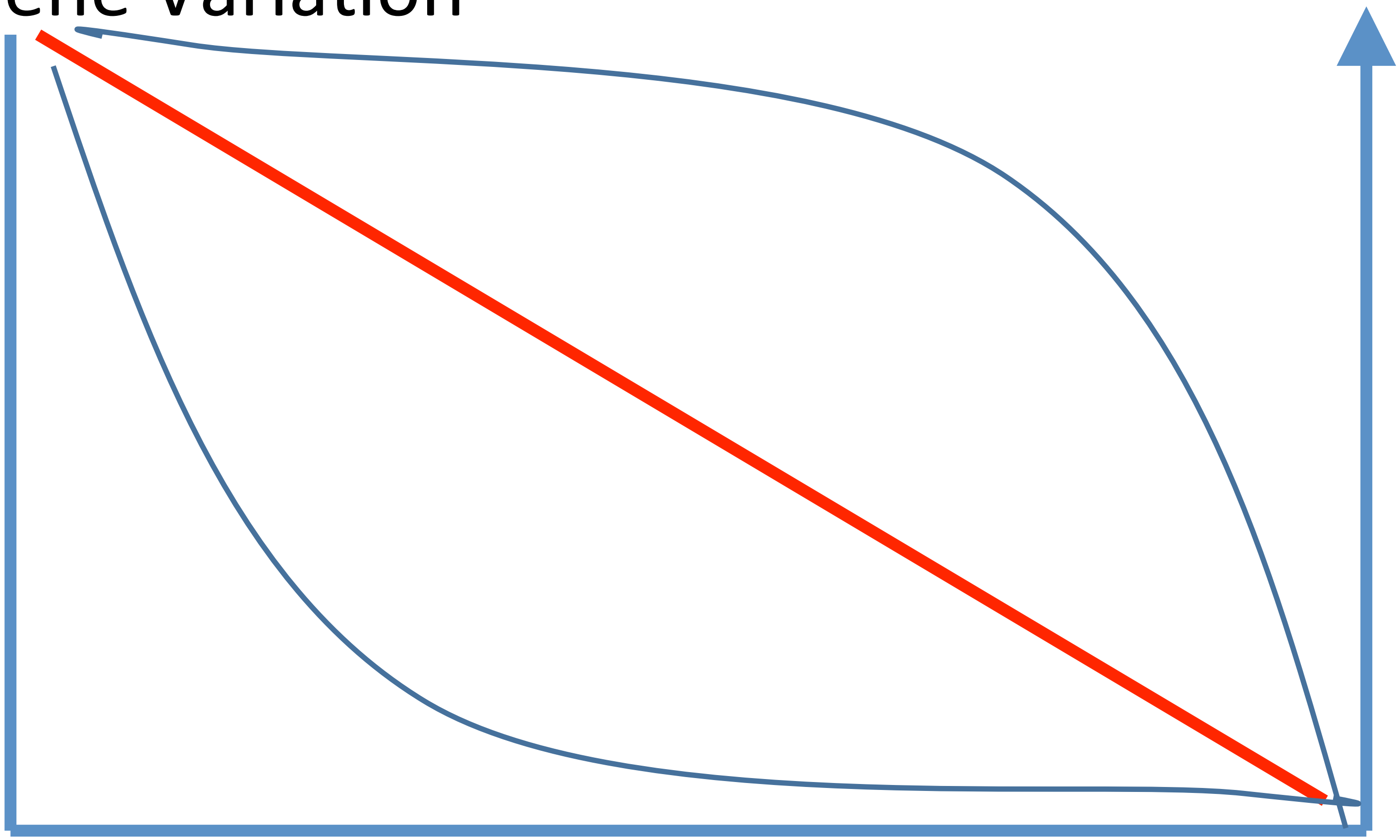
Geno + Enviro = Pheno (GEP graph)



Jong Bhak. Under BioLicense: public domain

Single Gene . Environe . Phene Variation

- Gene Variation



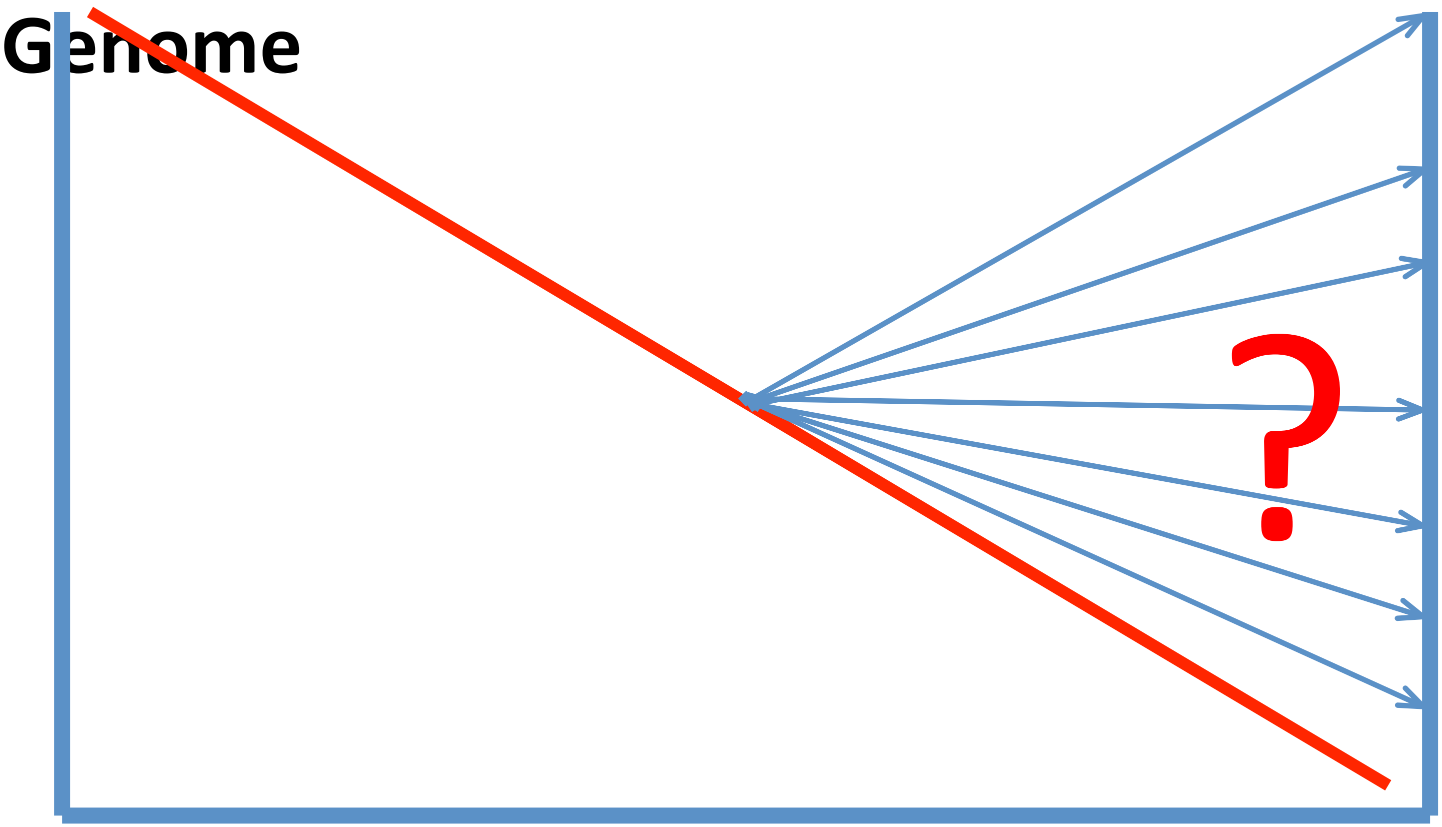
Genome Envirome and Phenome

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

GenoEnviroPheno Unpredictability Graph

Phenome

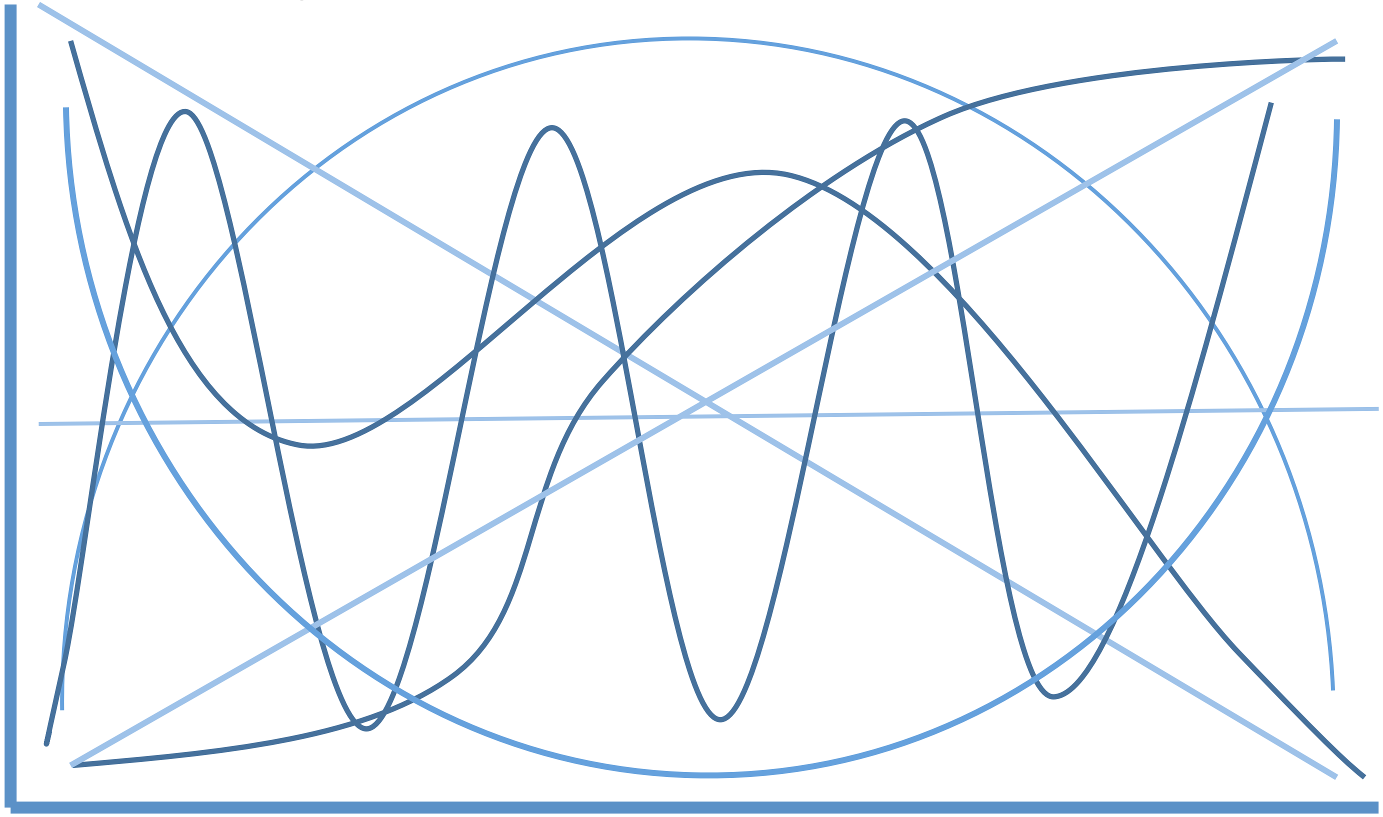
- Genome



Envirome

Gene-complex ←→ Phene-complex

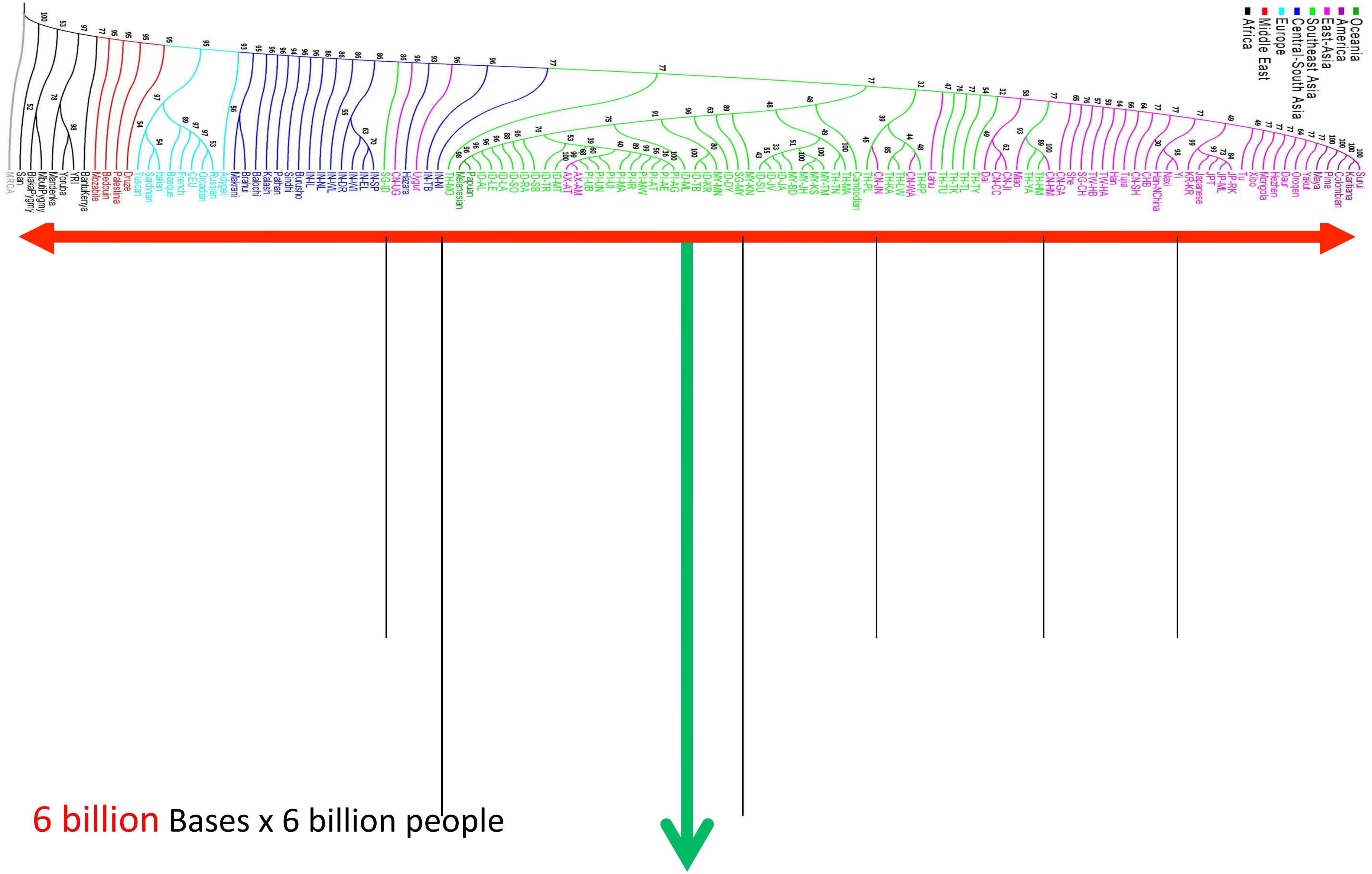
- GeneComplex



PheneComplex

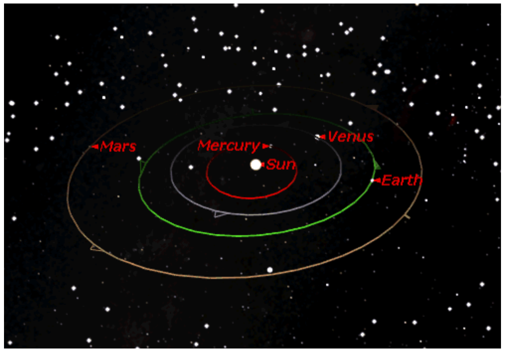
Jong Bhak. Under BioLicense: public domain

We must Find Structure in Population Matrix



6 billion Bases x 6 billion people

Genome size



- 0.00000000034 meter X 6 billion => 2.04 meter
  - – 2.04 x 6 billion => 12,240,000,000 meters
- Sun’s diameter: 1,380,000,000 → about 9 suns
- A long string, alignment, and phylogeny problem.

6 Billion Genomics

- 1 person($1,000) => 6,000,000,000,000 ($6 trillion)
- 3 GB x 6 GB => 1.8e+19 DNA base pairs
  - Reading it 40x ➔ 7.2e+20 base pairs
  - **2.4 billion 3TB HDDs**
  - 100 GB (1 person)
    - ➔ takes 1 week to get useful BAM, VCF files using 250 core 512GB, 32GB individual board memory Cluster
    - ➔ 420 billlion weeks ➔ 807,692,307 years
    - ➔ GPU ➔ can be one in one day instead of 7 days
      - 115,384,615 years
    - ➔ Energy: Running 20kw (1 kw = $0.07064 ➔ $1,400)
      - Approximately $365,000 per year. (discount rate ☺)
      - $52,115,384,475,000 (52 trillion USD ← 6 billion x $365,000)

Individual Variome

- Each person has about 4 million SNP (small size variants)
  - => 6 billion X 4 mil ➔ 2.4e+16 variants
  - Cancer samples ➔ every year 7 million people die of cancer ➔ 70 million cancer patients.
  - Each cancer genome is its own species ➔ 2.1e+17 cancer variants
  - 2.4e+16 + 2.1e+17 variants to process
  - Analyzing one cancer genome takes at least weeks.
    - Extracting variants and comparing them (align) with DBs
    - Every single variant is usually not a sington ➔ network of variants interactions ➔ non-linear
    - If it takes ONE hour to process one cancer patient's total variants:
      - Analyzing 70,000,000 cancer genomes ➔ 7,990 years of computing

Benefits of applying innovative algorithms

- Compression
- **Efficient Difference comparison**
- **Standardization**
- Parallelization: Hardware & Software (MIC)
- Automation
- Ease of use
- Visualization for lay people

Suggestion

- Big Genomic Data Programming & Infrastructure researching on:
  - Cost-Efficiency in pipelines
  - Standardization (taking up users' needs quickly)
  - Efficient core algorithms
  - Databases (cheap and fast enough)
- Not another authority or bureaucracy
  - Virtual Institute or Consortium

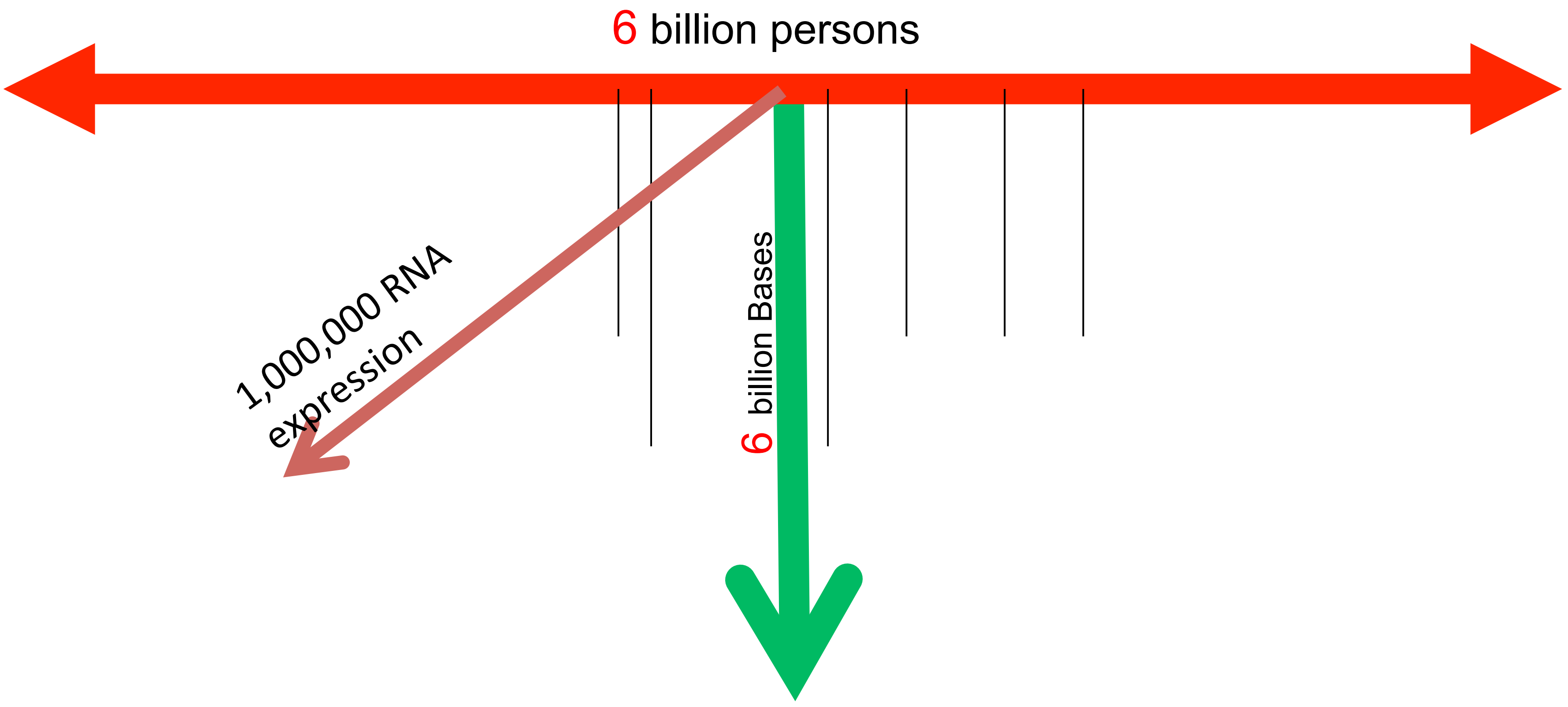
Increase

- 2012 ➔ 10,000s human genomes sequenced ➔ The rate is ~10x per year.

- Not only that….

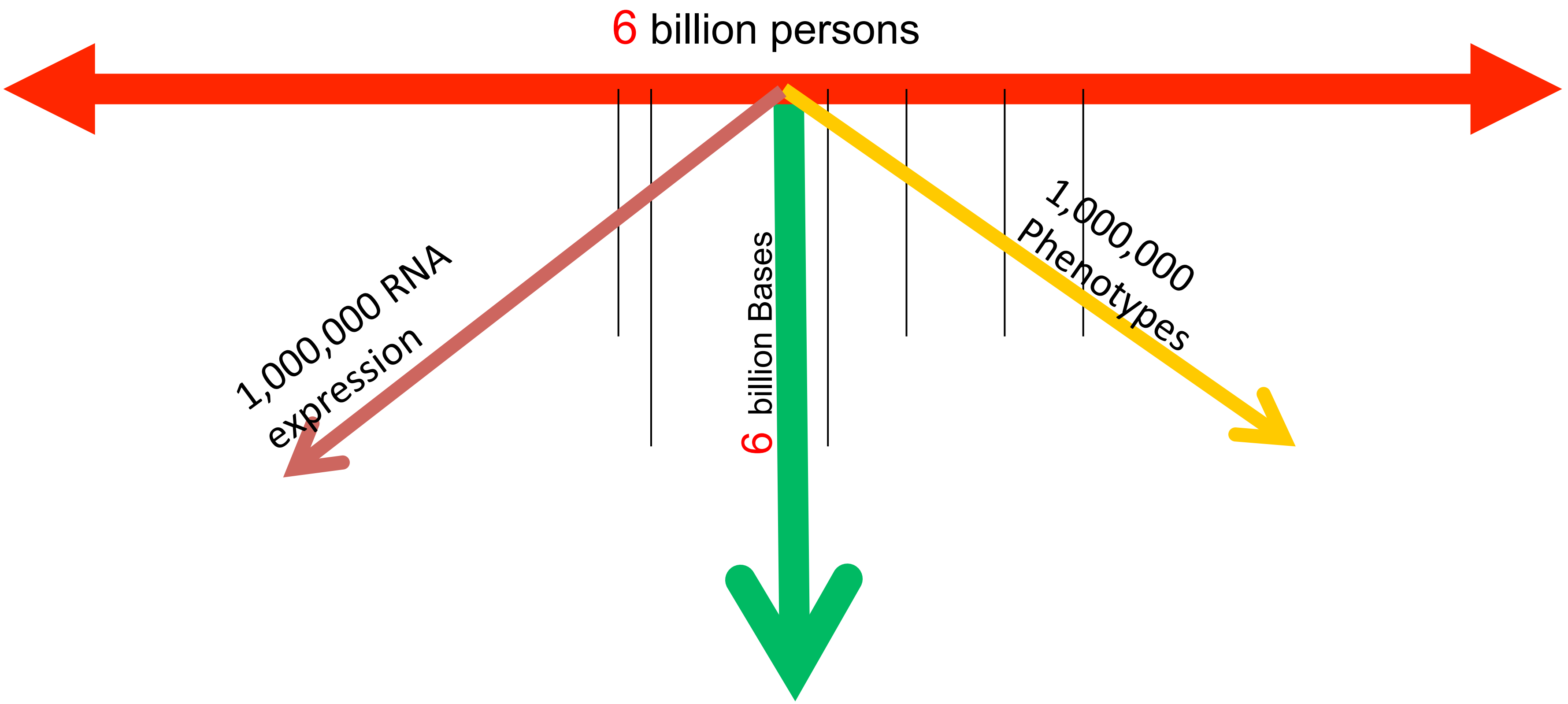
Adding one more dimension?

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

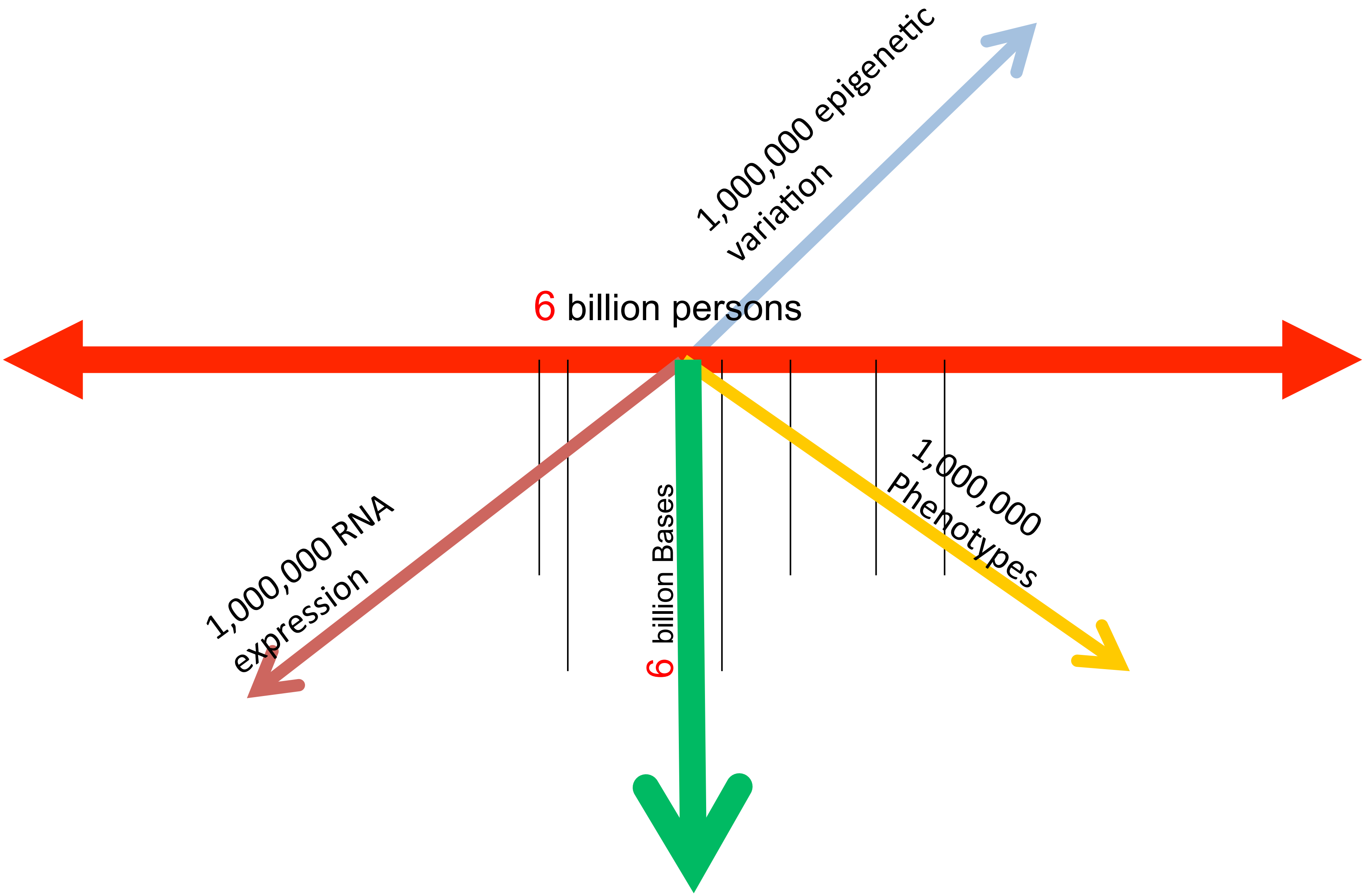


Adding even more dimensions?

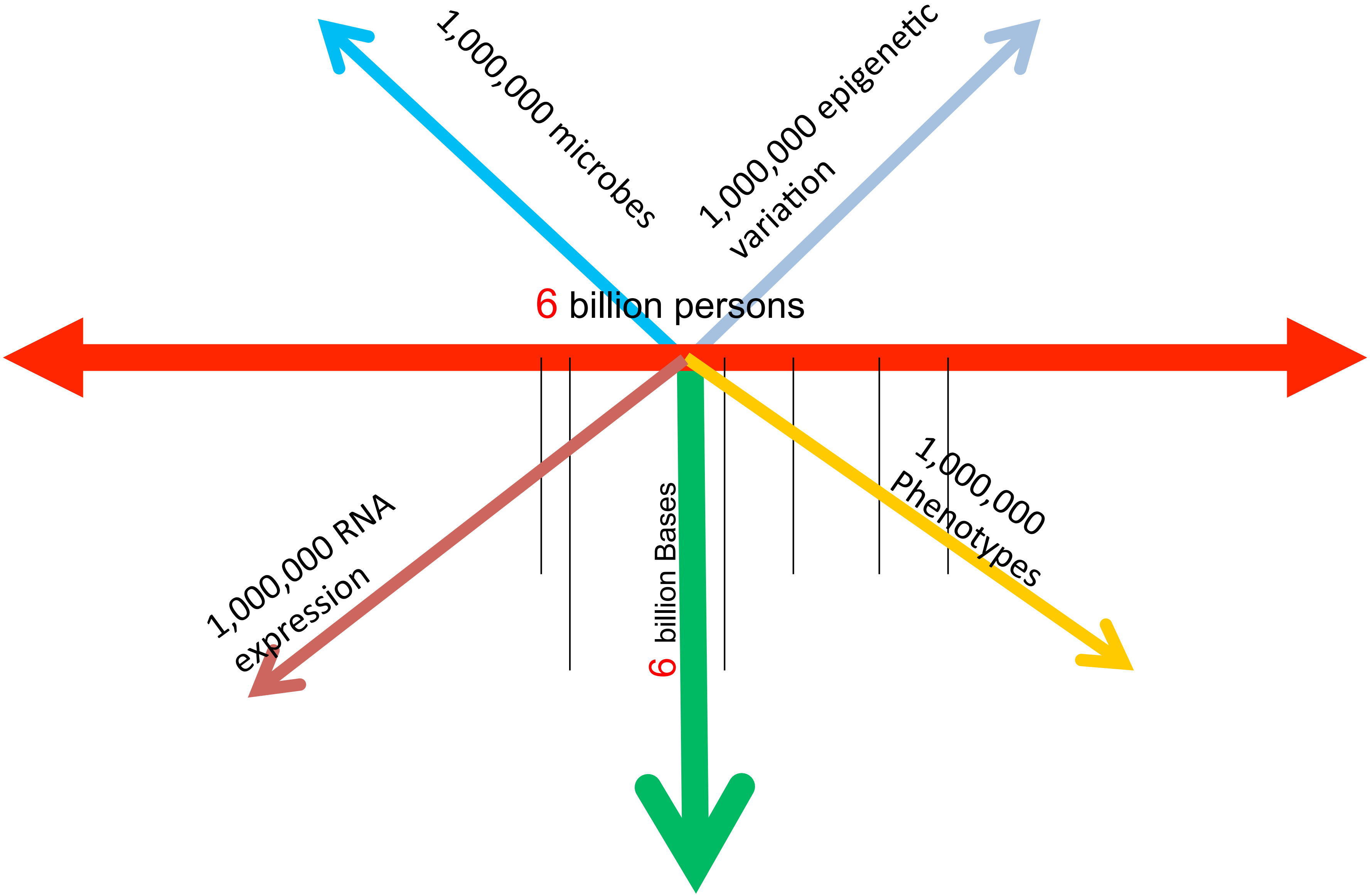
How to map/compute **Phenome?**



How to map/compute **epigenome**?



How to map/compute **Microbiome**?



KOREAN PERSONAL GENOME PROJECT

(KPGP)

Personal Genome Project (PGP)

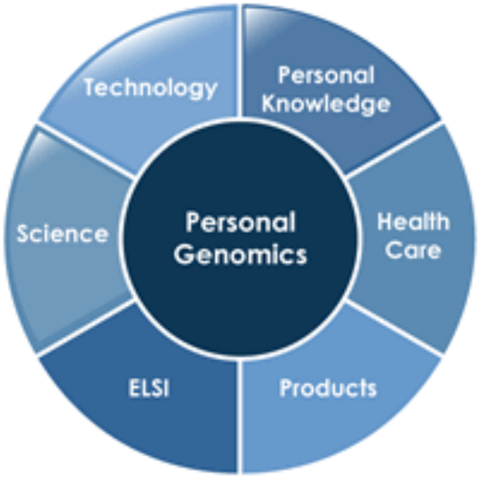
➢ Public Open Source Genome Project

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

➢ Led by Prof. **George Church** at 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:

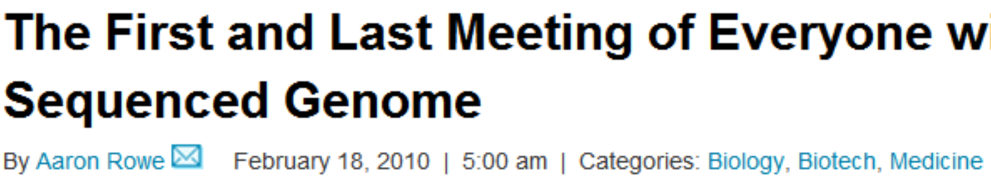
➢ are effective, informative, and responsible

➢ yield identifiable and improvable benefits at manageable levels of risk

➢ are broadly available for the good of the general public

The GET Conference 2010 brought together more than a dozen genome pioneers on the same stage to share their experiences and discuss the important ways in which personal genomes will affect all of our lives in the coming years. The conference was held April 27, 2010 in Cambridge, MA.

GET Genomes Environments Traits CONFERENCE



The First and Last Meeting of Everyone with a Fully Sequenced Genome

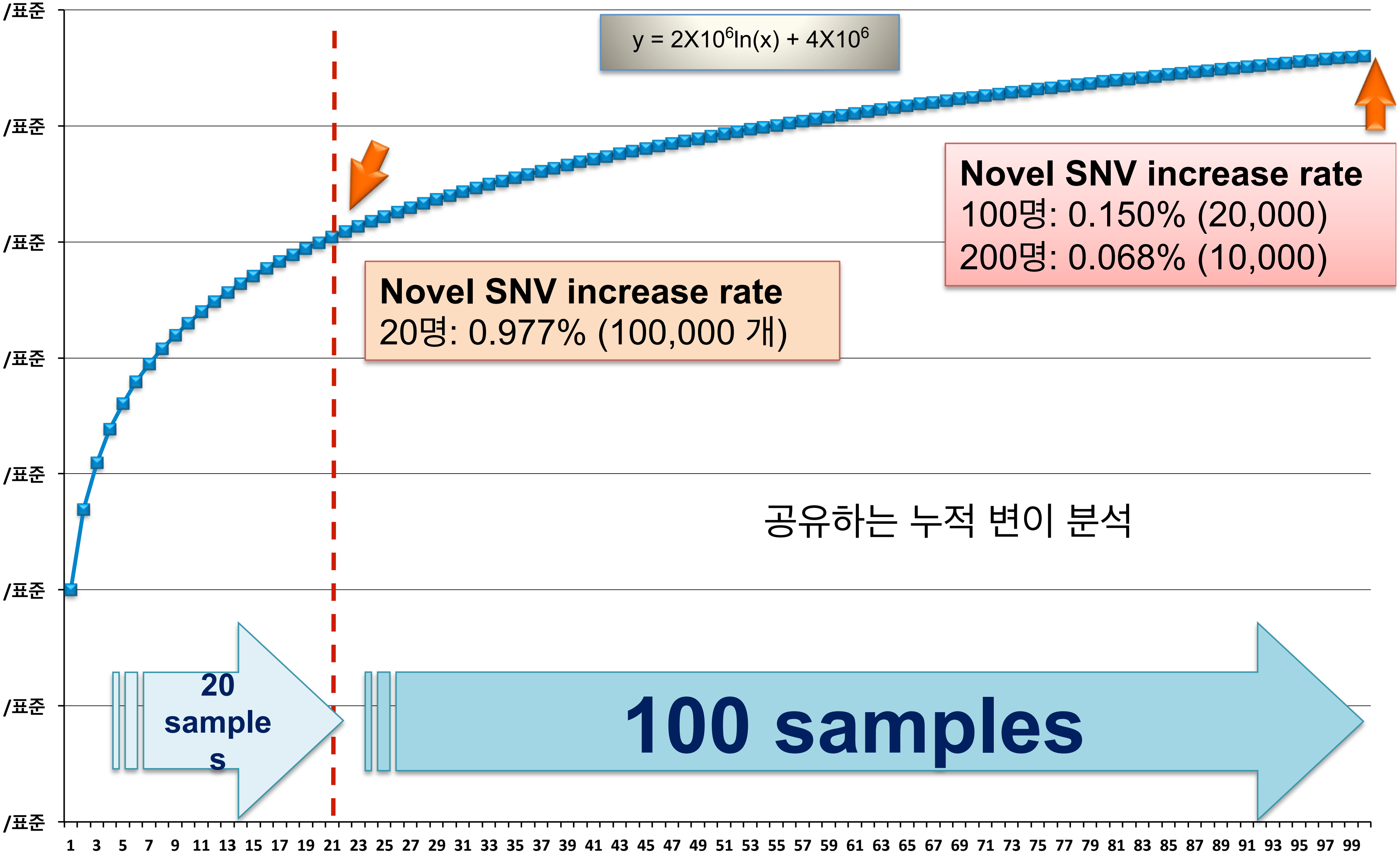
By Aaron Rowe ✉ February 18, 2010 | 5:00 am | Categories: Biology, Biotech, Medicine

- Extension of Harvard PGP Project in Korea

- Period : 2007 -2022

- Plan
  - **1단계 2007 ~ 2009, 1**
  - **2단계 2010 ~ 2011, 100**
  - **3단계 2012 ~ 2013, 3,000**
  - **4단계 2014 ~ 2017, 10,000**
  - **5단계 2017 ~ 2022, 50,000,000**

KPGP-20 Results

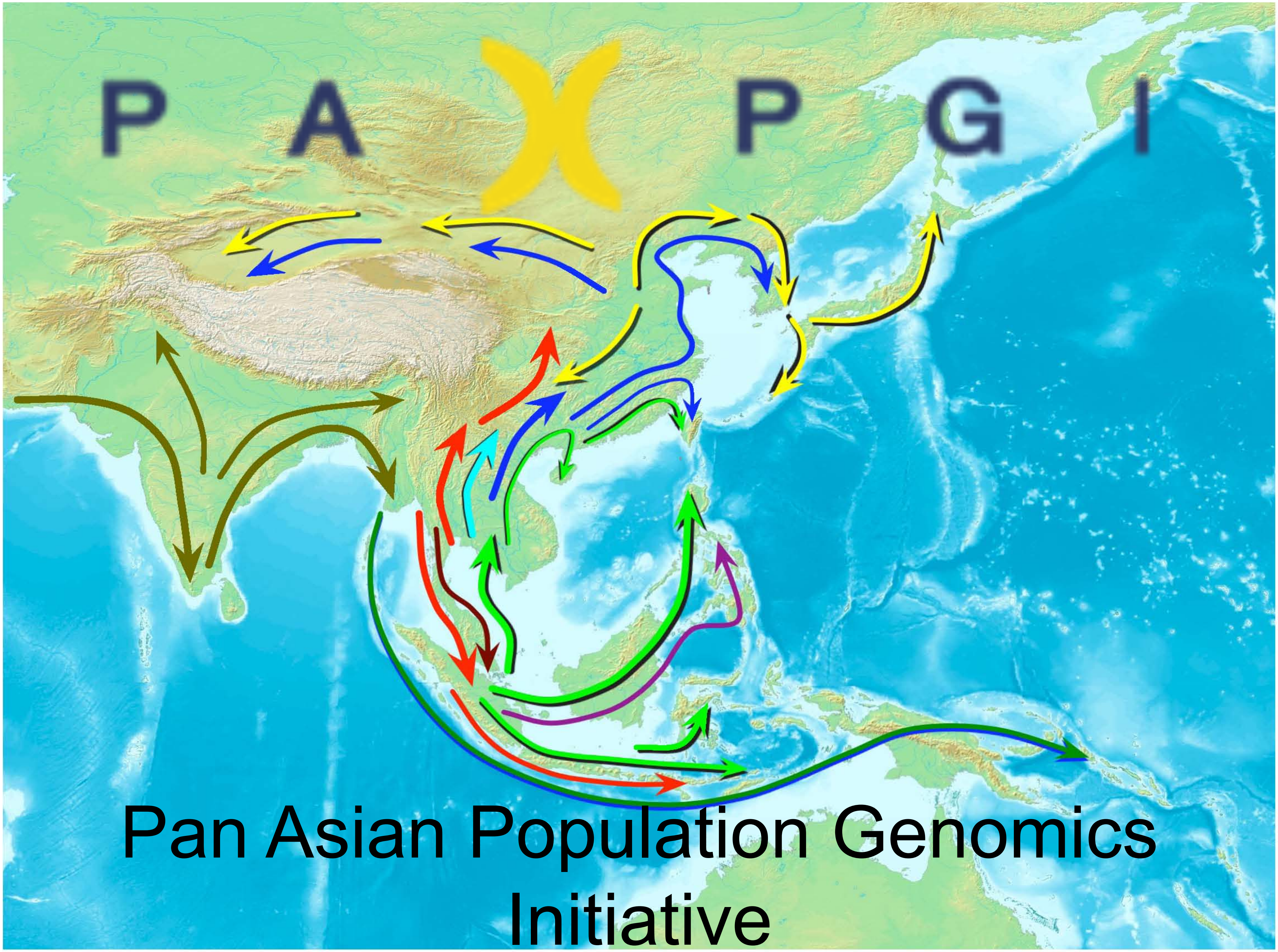


$y = 2X10^6 \ln(x) + 4X10^6$

Novel SNV increase rate
20명: 0.977% (100,000 개)

Novel SNV increase rate
100명: 0.150% (20,000)
200명: 0.068% (10,000)

공유하는 누적 변이 분석



Pan Asian Population Genomics Initiative

Introducing PASNP

- **Pan Asian SNP initiative**

  (PASNP 1.0)

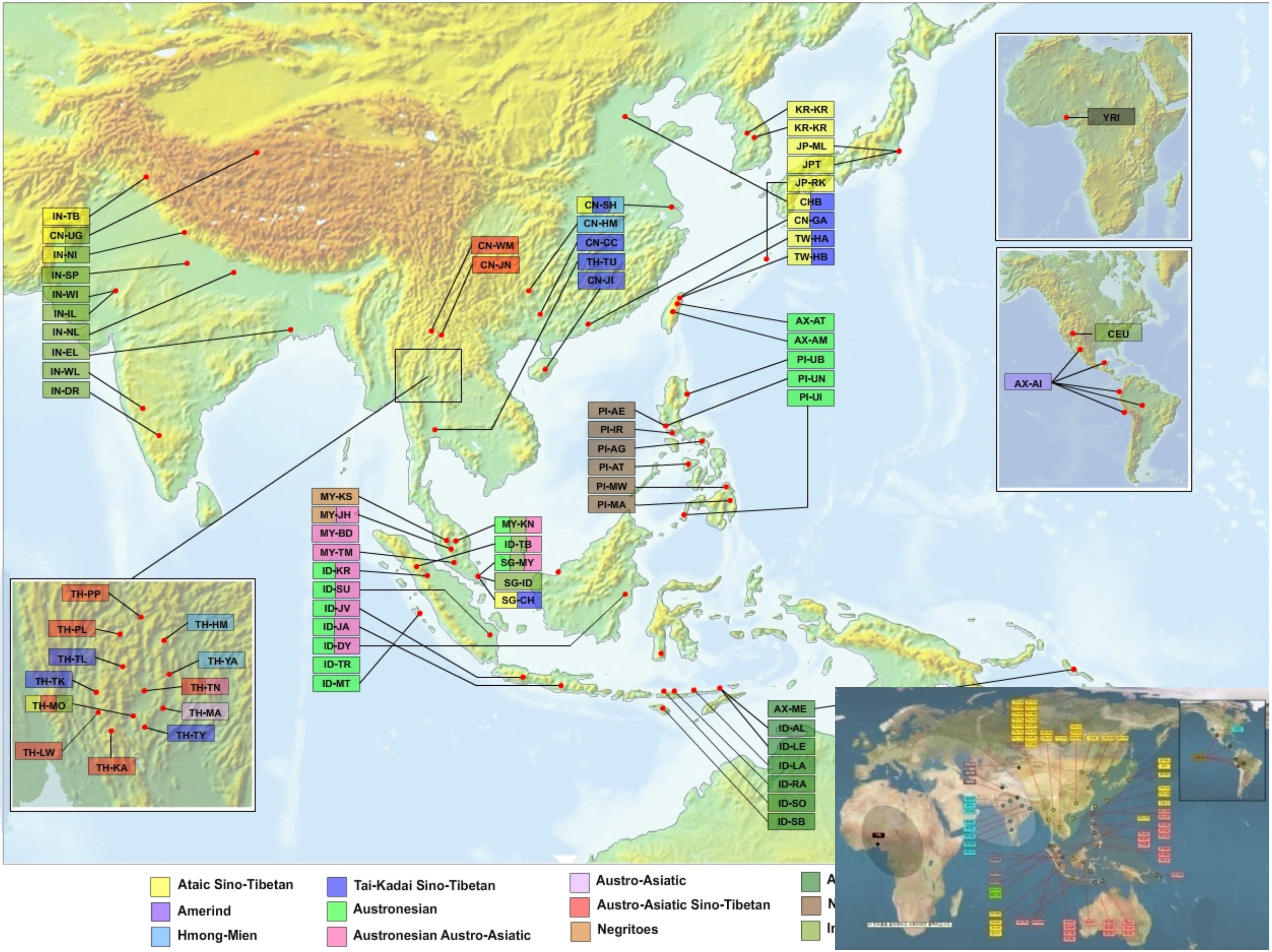
http://pasnp.net

http://papgi.org

Samples from 11 Pan Asian countries

Sample number: ~2000
Ethnic group: 76
Country: 11
SNP marker number: 58,960

(Affymetrix 56K Xba SNP genotyping chip)


IN-TB
CN-UG
IN-NI
IN-SP
IN-WI
IN-IL
IN-NL
IN-EL
IN-WL
IN-DR
CN-WM
CN-JN
CN-SH
CN-HM
CN-CC
TH-TU
CN-JI
KR-KR
KR-KR
JP-ML
JPT
JP-RK
CHB
CN-GA
TW-HA
TW-HB
AX-AT
AX-AM
PI-UB
PI-UN
PI-UI
PI-AE
PI-IR
PI-AG
PI-AT
PI-MW
PI-MA
MY-KS
MY-JH
MY-BD
MY-TM
ID-KR
ID-SU
ID-JV
ID-JA
ID-DY
ID-TR
ID-MT
MY-KN
ID-TB
SG-MY
SG-ID
SG-CH
AX-ME
ID-AL
ID-LE
ID-LA
ID-RA
ID-SO
ID-SB
TH-PP
TH-PL
TH-TL
TH-TK
TH-MO
TH-LW
TH-KA
TH-HM
TH-YA
TH-TN
TH-MA
TH-TY
YRI
CEU
AX-AI
Ataic Sino-Tibetan
Amerind
Hmong-Mien
Tai-Kadai Sino-Tibetan
Austronesian
Austronesian Austro-Asiatic
Austro-Asiatic
Austro-Asiatic Sino-Tibetan
Negritoes
A
N
Ir

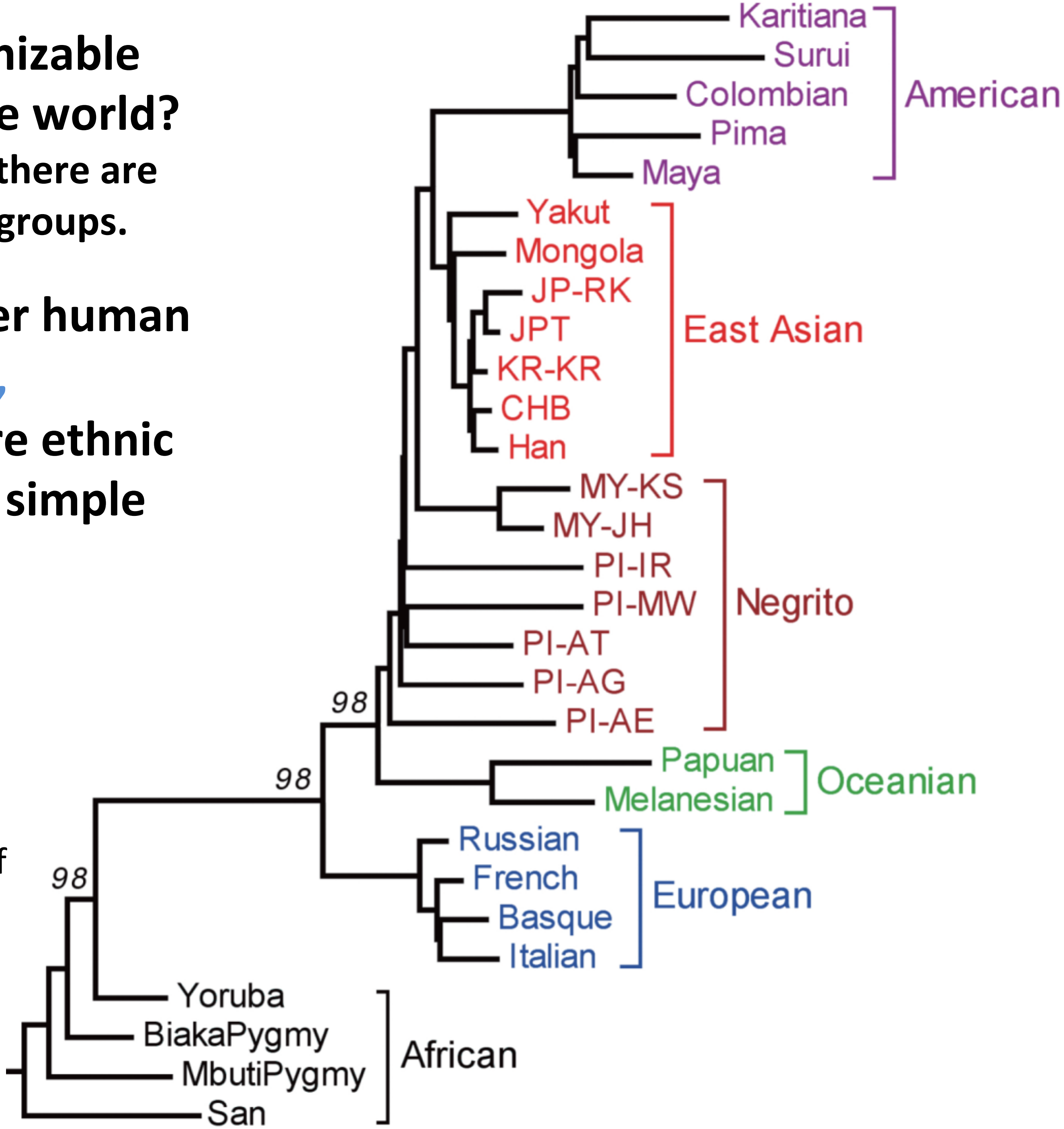
Genotyped 76 ethnic groups over 11 countries

Ethnic group code	Ethnicity	Ethnic group code	Ethnicity	Ethnic group code	Ethnicity
AX-AI	Karitiana, Maya, Quechua, Auca, Pima	ID-SU	Sunda	PI-MA	Minanubu
AX-AM	Ami	ID-TB	Batak Toba	PI-MW	Mamanwa
AX-AT	Atayal	ID-TR	Toraja	PI-UB	Filipino
AX-ME	Melanesians	IN-DR	Proto-Austroloids	PI-UI	Filipino
CEU	European	IN-EL	Caucasoids (may have admixture with Mongoloids)	PI-UN	Filipino
CHB	Han	IN-IL	Caucasoids	SG-CH	Chinese
CN-CC	Zhuang	IN-NI	Mongoloid features	SG-ID	Indian
CN-GA	Han	IN-NL	Caucasoids	SG-ML	Malay
CN-HM	Hmong	IN-SP	Caucasoids	TH-HM	Hmong (Miao)
CN-JI	Jiamao	IN-TB	Mongoloid features	TH-KA	Karen
CN-JN	Jinuo	IN-WI	Caucasoids	TH-LW	Lawa
CN-SH	Han	IN-WL	Caucasoids	TH-MA	Mlabri
CN-UG	Uyghur	JP-ML	Japanese	TH-MO	Mon
CN-WA	Wa	JP-RK	Ryukyuan	TH-PL	Paluang
ID-AL	Alorese	JPT	Japanese	TH-PP	Plang
ID-DY	Dayak	KR-KR	Koreans	TH-TK	Tai Khuen
ID-JA	Javanese	MY-BD	Bidayuh	TH-TL	Tai Lue
ID-JV	Javanese	MY-JH	Negrito	TH-TN	H'tin
ID-KR	Batak Karo	MY-KN	Malay	TH-TU	Tai Yuan
ID-LA	Lamaholot	MY-KS	Negrito	TH-TY	Tai Yong
ID-LE	Lembata	MY-MN	Malay	TH-YA	
ID-ML	Malay	MY-TM	Proto-Malay	TW-HA	Chinese
ID-MT	Mentawai	PI-AE	Ayta	TW-HB	Chinese
ID-RA	Manggarai	PI-AG	Agta	YRI	Yoruba
ID-SB	Kambera	PI-AT	Ati		
ID-SO	Manggarai	PI-IR	Iraya		

- **How many recognizable human groups in the world?**

➔Just in the right fig., there are simply six recognizable groups.

➔When we consider human migration, isolation, admixture, and more ethnic groups, this is not a simple question.

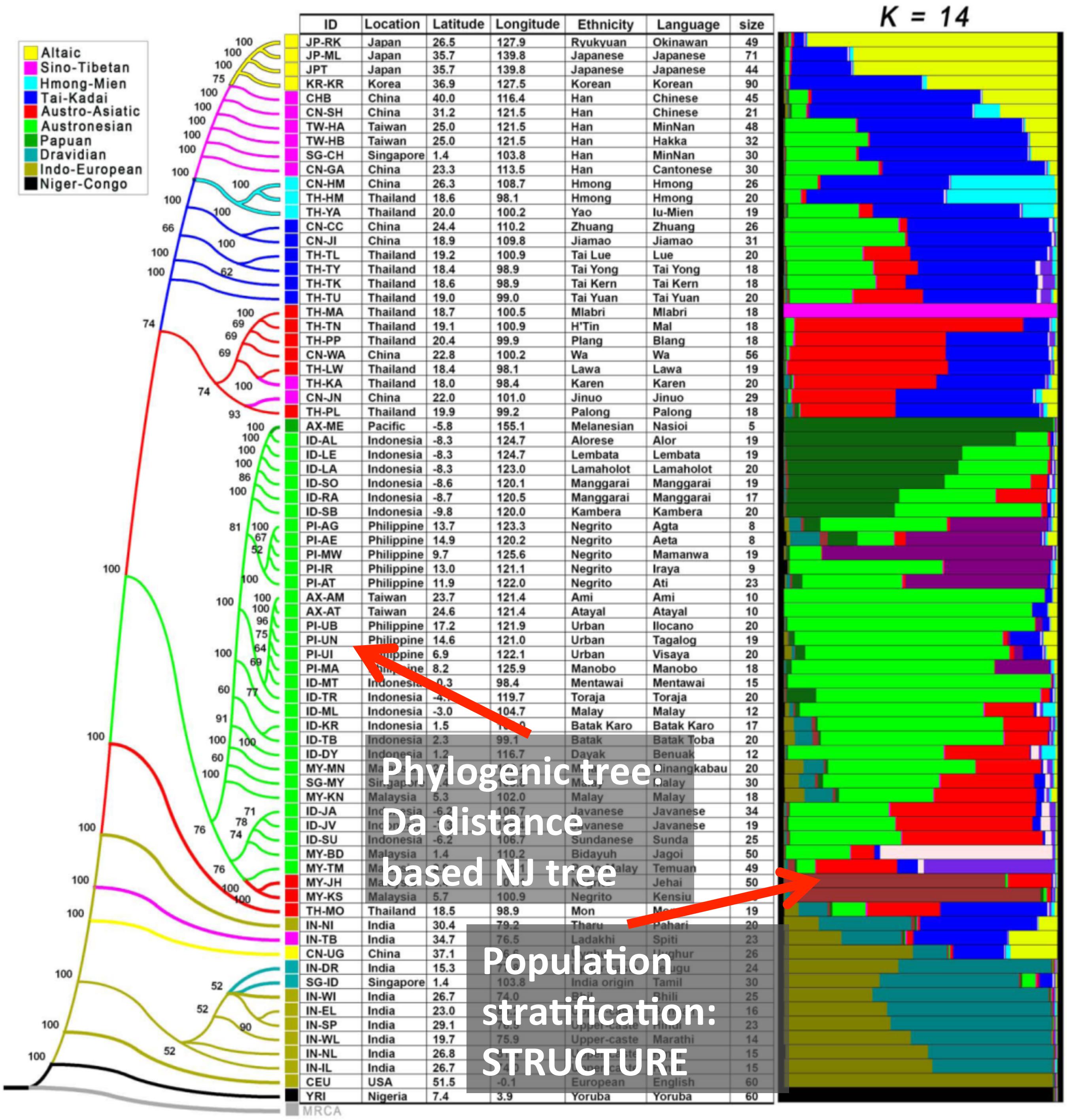


Maximum likelihood tree of 29 populations. The tree based on 19,934 SNPs. Bootstrap values based on 100 replicates

Phylogentic and population structure analysis results

Finding 1: Genetic ancestry is strongly correlated with linguistic affiliations, as well as geography.

Finding 2: Most populations show relatedness within ethnic/linguistic groups despite prevalent gene flow amongst populations.



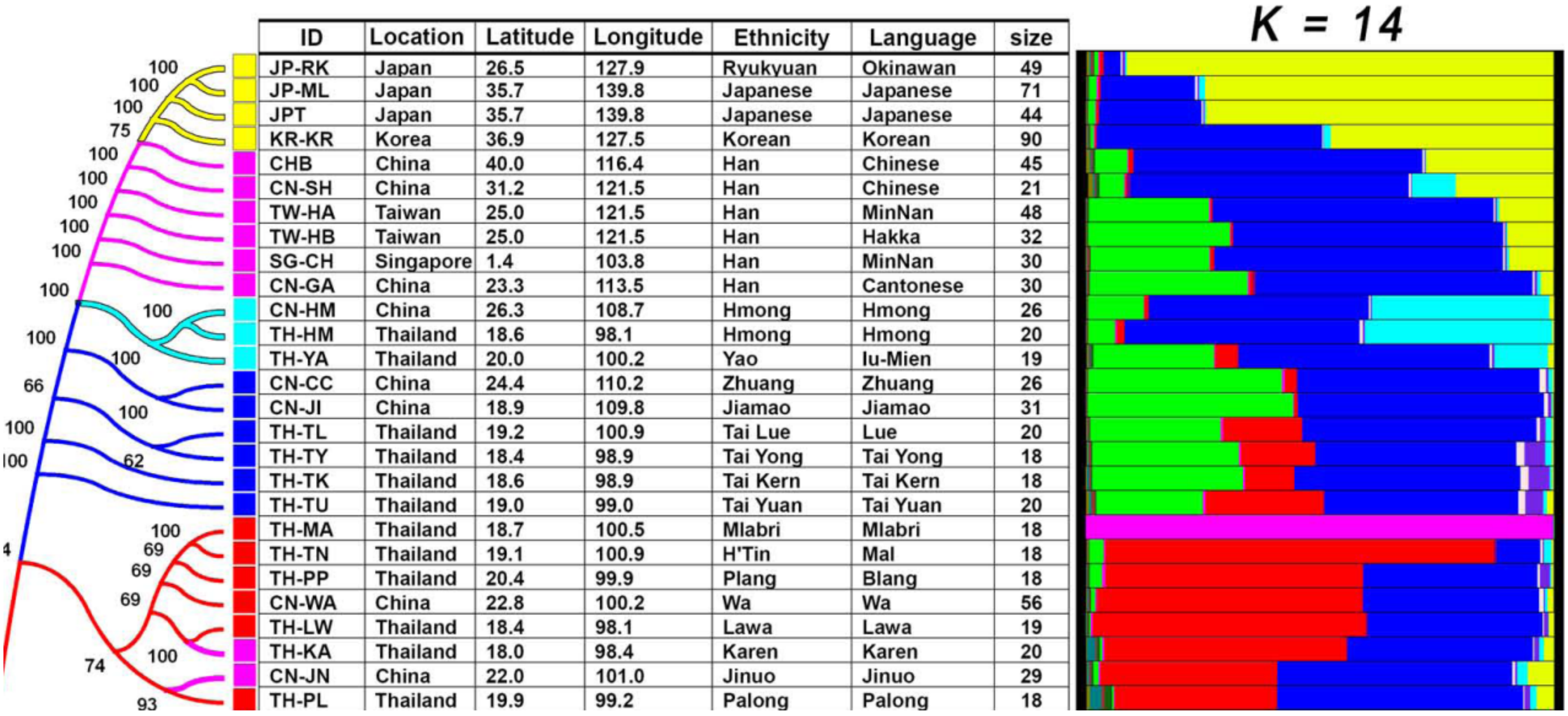
ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippine	13.7	123.3	Negrito	Agta	8
PI-AE	Philippine	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippine	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippine	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippine	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippine	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippine	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippine	6.9	122.1	Urban	Visaya	20
PI-MA	Philippine	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	[illegible]	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	[illegible]	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	[illegible]	[illegible]	[illegible]	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	[illegible]	[illegible]	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	[illegible]	[illegible]	Proto-Malay	Temuan	49
MY-JH	Malaysia	[illegible]	[illegible]	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	[illegible]
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	[illegible]	Uyghur	Uyghur	26
IN-DR	India	15.3	[illegible]	[illegible]	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	[illegible]	Upper-caste	[illegible]	16
IN-SP	India	29.1	[illegible]	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	[illegible]	Upper-caste	[illegible]	15
IN-IL	India	26.7	[illegible]	Upper-caste	[illegible]	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

Phylogenic tree: Da distance based NJ tree

Population stratification: STRUCTURE

Considerable gene flow among Asian populations was observed

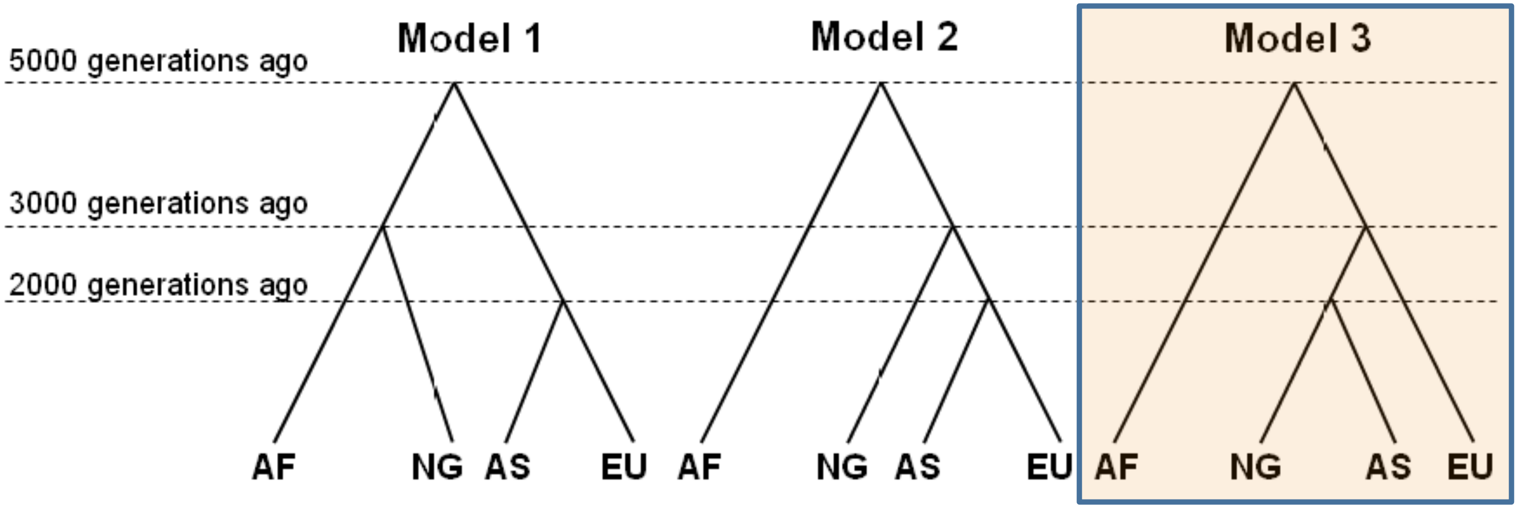
- Considerable gene flow was observed amongst sub-populations in clusters, including those groups believed to practice endogamy based on linguistic, cultural and ethnic information.



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18

Results and Conclusion:

Peopling of Asia: one-wave versus two-wave hypothesis



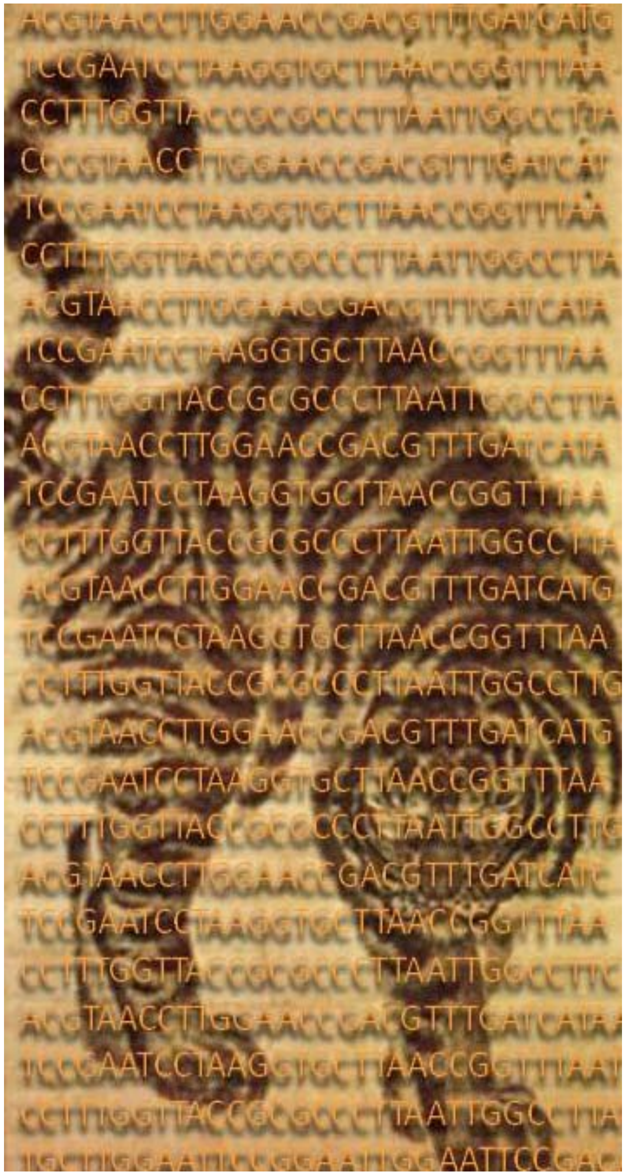
- Our simulation results indicate that Model 1 is not compatible with the empirical data,
- Model 2 is the only compatible if gene flow from other Asian populations to the Negritos has been fairly extreme, with more than 50% of Negrito chromosomes coming from other Asian populations, without dramatically affecting the Negrito phenotype.
- **Thus Model 1 and 2 are impertinent to the explanation of current observations.**

No extreme gene flow!

People of Thailand

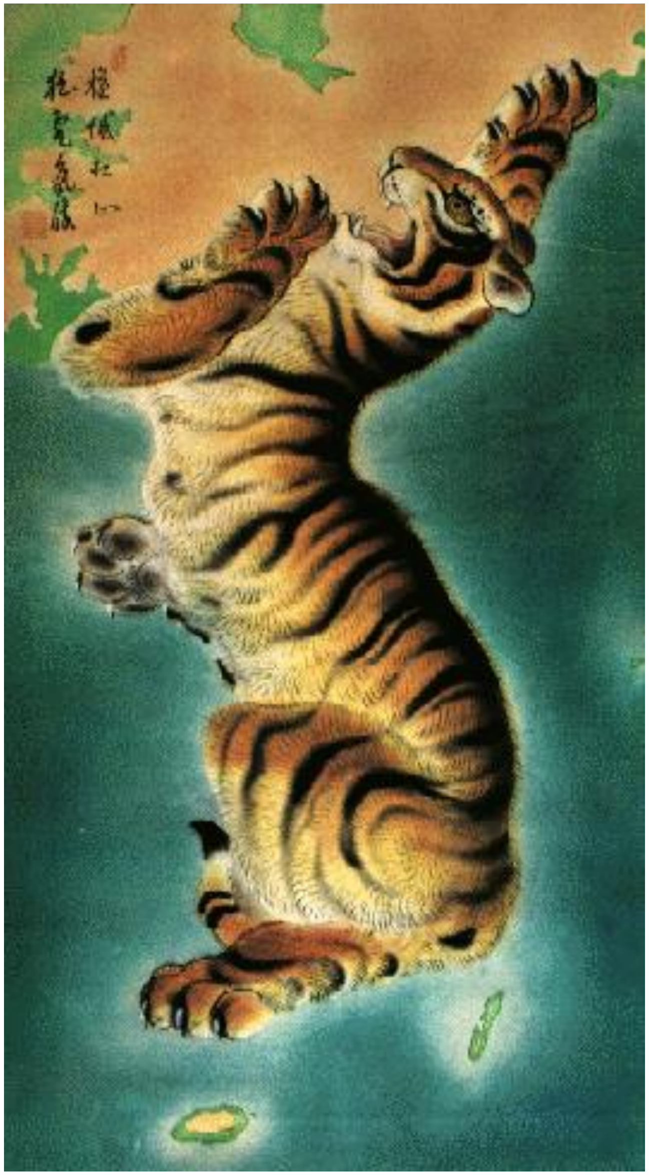
Negrito: The Semang people of the Malay Peninsula

Open Tiger Genome Project



PGI, GRF, TBI
BGI, SNU, SSU, …

http://tigergenome.org



TaeGeuk (Amur tiger)

HwaRang (White tiger)

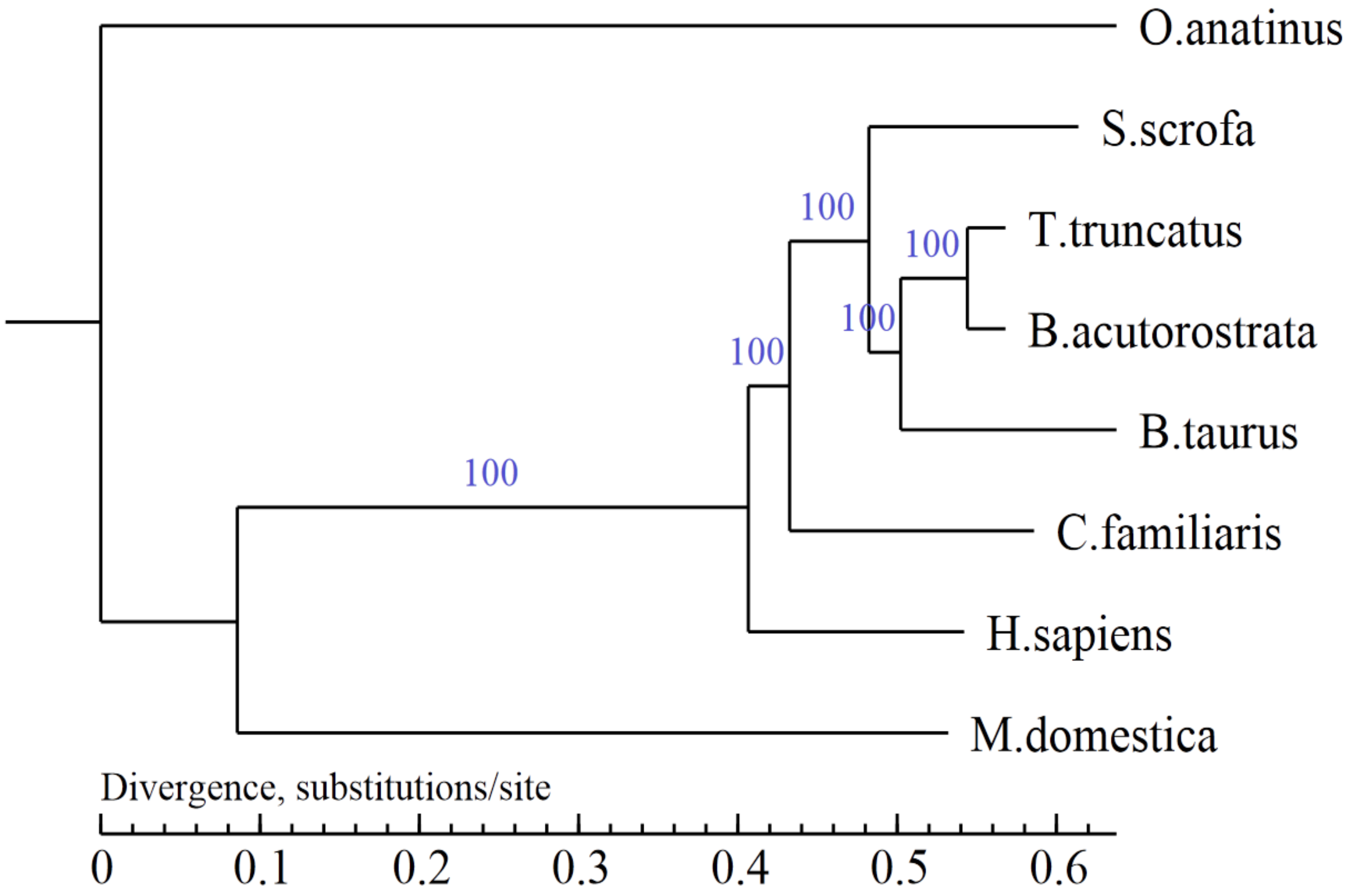
SunDol (African lion)

SnowGirl (White lion)

Whale Genome Project

- KIOST and TBI
- Minke whale (*Balaenoptera acutorostrata)*
- 2.8 GB
- Over 200 GB data

- http://whalegenome.net


O.anatinus
S.scrofa
100
T.truncatus
100
100
B.acutorostrata
100
B.taurus
100
C.familiaris
H.sapiens
M.domestica
Divergence, substitutions/site
0
0.1
0.2
0.3
0.4
0.5
0.6