BiG Genomics
(Billion Genome Project)

Jong Bhak
Theragen BiO Institute
San Diego
Acknowledgement

• 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 copyleft.
• 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”
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 live 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
- BioProgramming..

http://bioprogramming.org
http://biolinux.net
http://bioperl.net
http://biophp.net
http://bioos.org
http://biojava.net
BioEngine: Automatic BioInformation Pipelines Processing System
GiSys

Cloud GUI | Integrated DBs | Shopping cart and charge | Users portal

LIMS (Laboratory Information Management System)

Application Level

Bioinformatics Pipelines

Omics integrated DBs

Workflow based Middleware

Cloud based infrastructure

Cloud GUI | Integrated DBs | Shopping cart and charge | Users portal

LIMS (Laboratory Information Management System)

Application Level

Bioinformatics Pipelines

Omics integrated DBs

Workflow based Middleware

Cloud based infrastructure
To do What?
Geno + Enviro = Pheno (GEP graph)

- Geno
- Traits (pheno)
  - Each Trait or Disease
  - Cancer
  - Flu
  - Car Accident

Jong Bhak. Under BioLicense: public domain
Single Gene . Environ . Phene Variation

• Gene Variation

Environ Variation

Jong Bhak. Under BioLicense: public domain
Genome Envirome and Phenome

• Genome = gene types + their variome
• Envirome = environe types + their variome
• Phenome = phene types + their variome
GenoEnviroPheno Unpredictability Graph

- Genome

Envirome

Jong Bhak. Under BioLicense: public domain
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

• 0.000000000034 meter × 6 billion ⇒ 2.04 meter
  – 2.04 × 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 billion 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 $\Rightarrow$ 10,000s human genomes sequenced
  $\Rightarrow$ The rate is $\sim$10x per year.

- Not only that....
Adding one more dimension?

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

6 billion Bases

6 billion persons

1,000,000 RNA expression
Adding even more dimensions?

How to map/compute Phenome?
How to map/compute epigenome?
How to map/compute Microbiome?

1,000,000 microbes
1,000,000 epigenetic variation
6 billion persons
1,000,000 RNA expression
6 billion Bases
1,000,000 Phenotypes
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.
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

Novel SNV increase rate
20 sample: 0.977% (100,000 개)
100 sample: 0.150% (20,000)
200 sample: 0.068% (10,000)

y = 2 \times 10^6 \ln(x) + 4 \times 10^6

공유하는 누적 변이 분석

100 samples
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)
# Genotyped 76 ethnic groups over 11 countries

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