국립보건연구원
National Institute of Health

한국인칩 컨텐츠 특징 및 성능 소개

문 상 훈 (국립보건연구원 유전체센터 유전체연구과)

단일염기다형성

Human Genome Project (1990~2003)

- total of $3 billion over this period

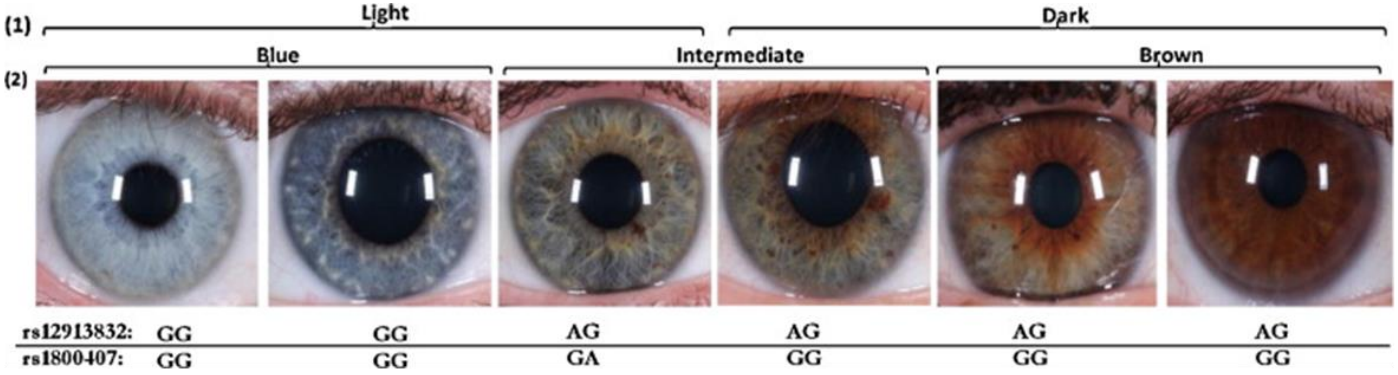
rs1234

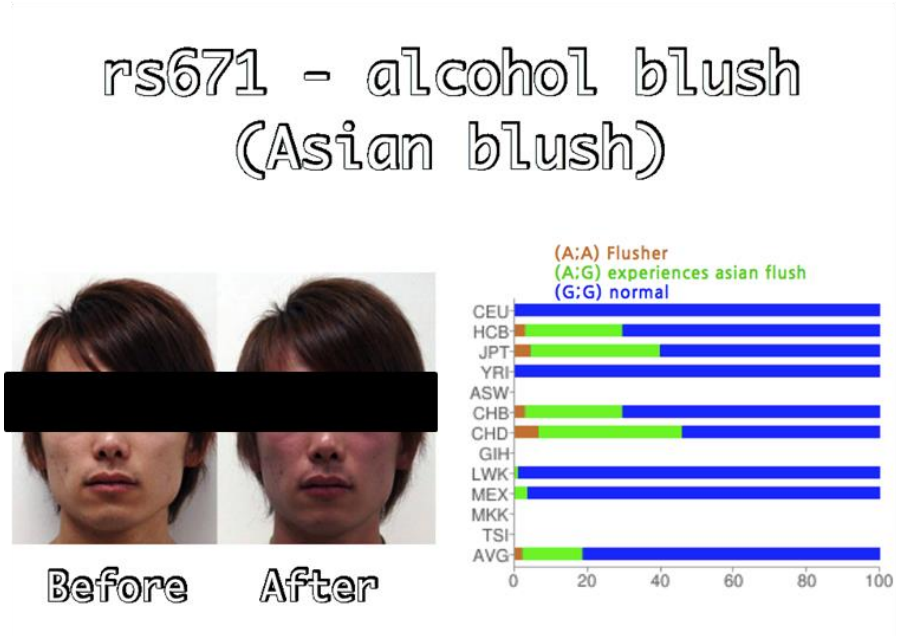
HGP	ATCTAAGGCTATGAGGGCCCAATCTAGCTTTTATTTCCGAGAGA
토마스	ATCTAAGGCTATGAGGGCCCAATCTAGCTTTTATTTCCGACAGA
제임스	ATCTAAGGCTATGAGGGCCCAATCAAGCTTTTATTTCCGACAGA
제인	ATCTAGGGCTATGAGGGCCCAATCAAGCTTTTATTTCCGACAGA

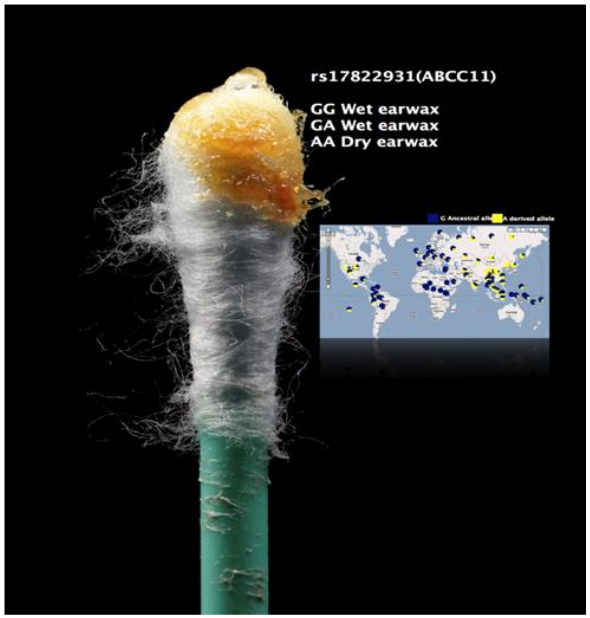
300~500bp

Single Nucleotide Polymorphism (SNP)

표현형의 차이


(1)
Light
Dark
(2)
Blue
Intermediate
Brown
rs12913832: GG GG AG AG AG AG
rs1800407: GG GG GA GG GG GG


rs671 - alcohol blush
(Asian blush)
Before
After
(A;A) Flusher
(A;G) experiences asian flush
(G;G) normal
CEU
HCB
JPT
YRI
ASW
CHB
CHD
GIH
LWK
MEX
MKK
TSI
AVG
0
20
40
60
80
100


rs17822931(ABCC11)
GG Wet earwax
GA Wet earwax
AA Dry earwax

2016. 7. 20. 서울신문

서울신문 www.seoul.co.kr since 1904

"한국인, 겨드랑이 냄새 덜 나는 유전자 가졌다" (연구)

한국인이 겨드랑이 냄새가 덜 나는 유전자를 가졌다는 연구 결과가 나왔다.

최근 영국의 브리스톨(Bristol) 대학은 영국인 6495 명을 조사한 결과 'ABCC11 유전자(ABC 수송체 유전자)'의 분포가 땀 냄새에 영향을 미친다고 밝혔다.

연구팀에 따르면 ABCC11 유전자는 'G 대립 유전자'와 'A 대립 유전자'로 나뉜다. 그 중 G유전자를 가진 사람은 겨드랑이 냄새를 유발하는 아포크린(Apocrine)땀샘의 땀 분비가 활발하게 이루어진다.

G유전자를 가질수록 땀 냄새가 많이 나고, A유전자를 가질수록 덜 난다는 말이다. G유전자는 주로 아프리카나 유럽인에게 나타나고 A유전자는 동아시아인이 많이 가지고 있는 것으로 나타났다.

이안 데이(Ian Day)교수는 "단지 2%의 유럽인만이 이 A유전자를 가지고 있다. 대부분의 동아시아인들은 A유전자를 가지고 있다"고 설명했다.

한국인은 G유전자가 거의 없는 것으로 알려졌다. 겨드랑이 냄새가 가장 덜 나는 민족임이 과학적으로 입증된 셈이다.

Frequencies of ABCC11 allele c.538 (One nonsynonymous SNP 538G > A)[23]

Ethnic groups	Tribes or inhabitants	AA	GA	GG
Korean	Daegu city inhabitants	100%	0%	0%
Chinese	Northern and southern Han Chinese	80.8%	19.2%	0%
Mongolian	Khalkha tribe	75.9%	21.7%	2.4%
Japanese	Nagasaki people	69%	27.8%	3.2%
Thai	Central Thai in Bangkok	63.3%	20.4%	16.3%
Vietnamese	People from multiple regions	53.6%	39.2%	7.2%
Native American		30%	40%	30%
Filipino	Palawan	22.9%	47.9%	29.2%
Kazakh		20%	36.7	43.3%
Russian		4.5%	40.2%	55.3%
White Americans	From CEPH families with out the French and Venezuelans	1.2%	19.5%	79.3%
African	From various sub-Saharan nations	0%	8.3%	91.7%
African Americans		0%	0%	100%

Wikipedia

A형 : 사타구니나 겨드랑이 아래에 아포크린땀샘이 적고 마른 귀지를 갖는다.

G형 : 사타구니나 겨드랑이 아래 아포크린샘이 많고 마른 귀지를 갖는다. 아포크린샘은 단백질, 지방질, 당질, 암모니아, 피루브산, 노화색소, 철분 등을 포함한 약간의 점성이 있는 땀을 분비하고 이런 성분을 세균이 분해할 경우 암내가 난다.

동아시아인들 끼리도 유전형이 다름

국립보건연구원
National Institute of Health

PCA analysis of East Asian descent

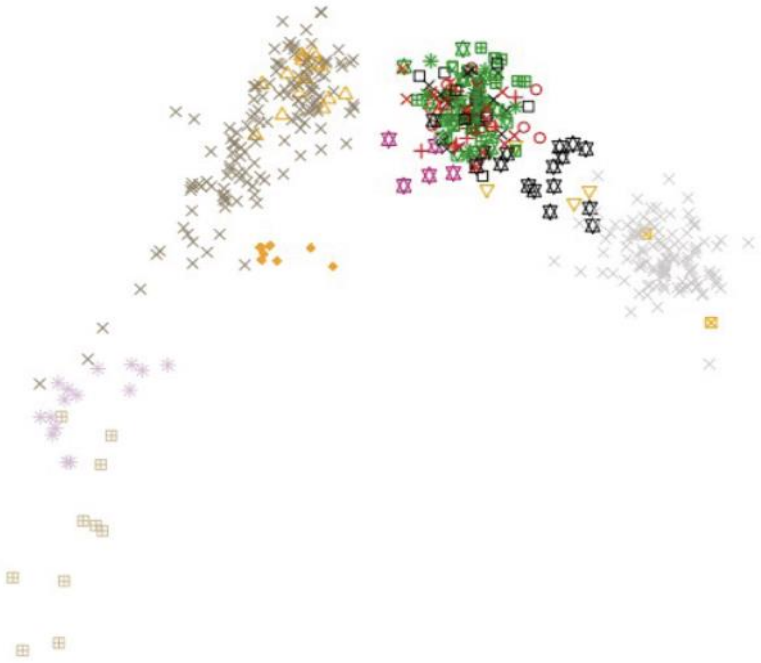
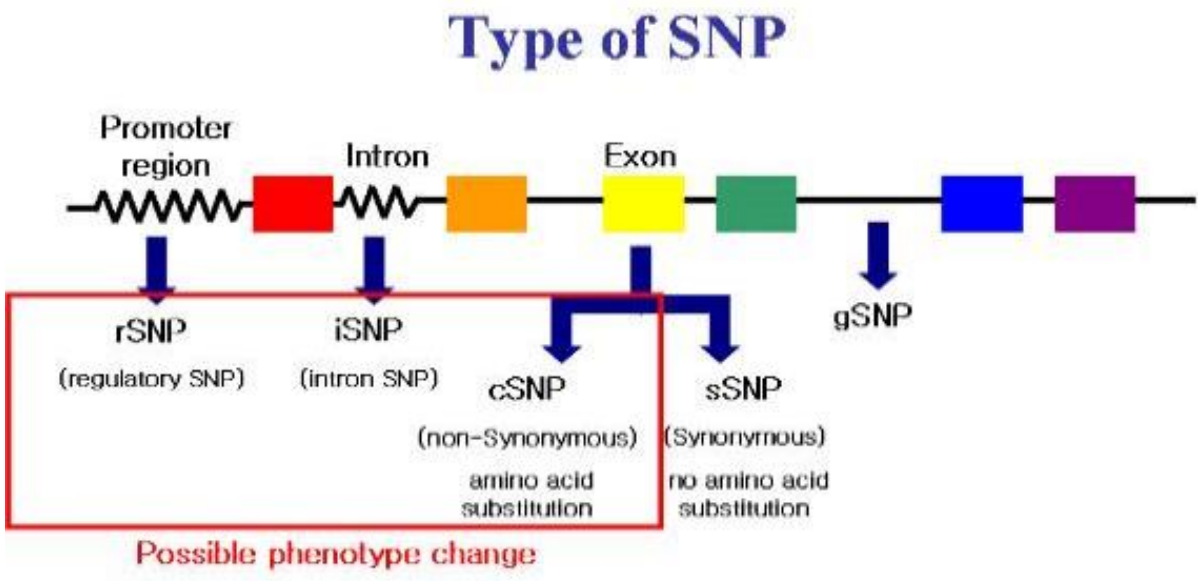


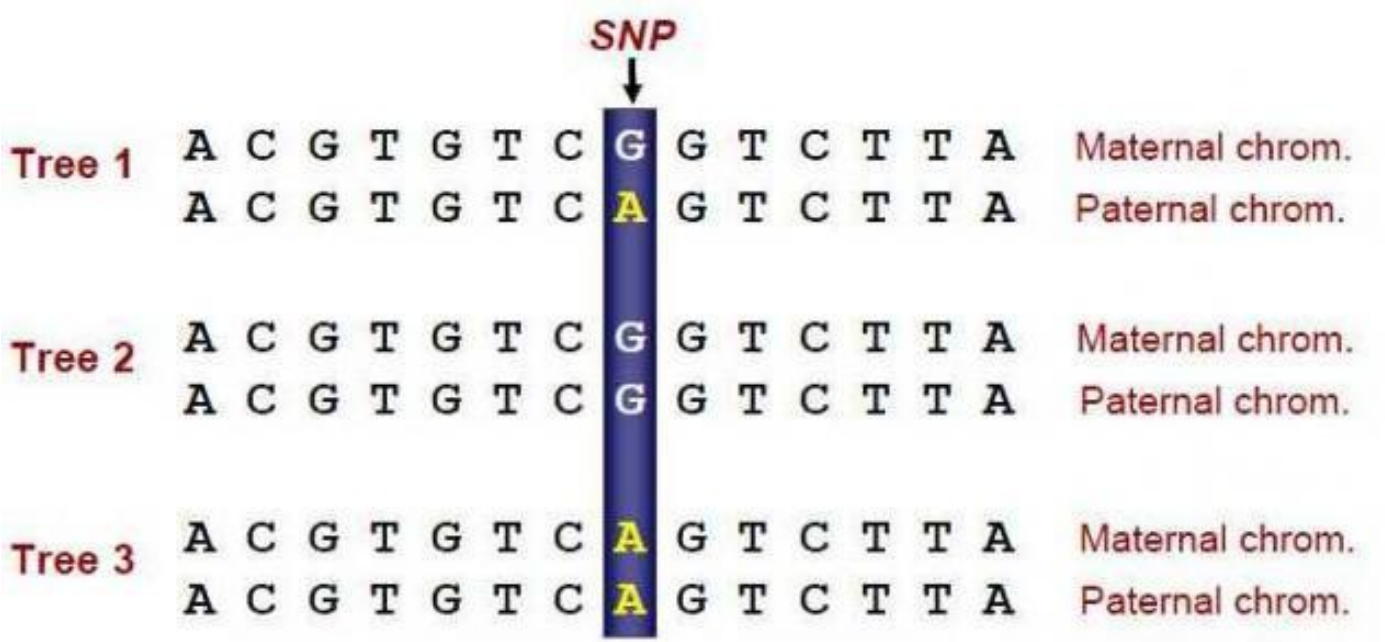
illustration of geographic correspondence of ethnic group locations

- Synonymous: do not result in a change of amino acid in the protein, but still can affect its function in other ways
- Non-synonymous
  - Missense : amino acid changes
  - Nonsense : changes amino acid to stop codon



Heterozygous vs. Homozygous

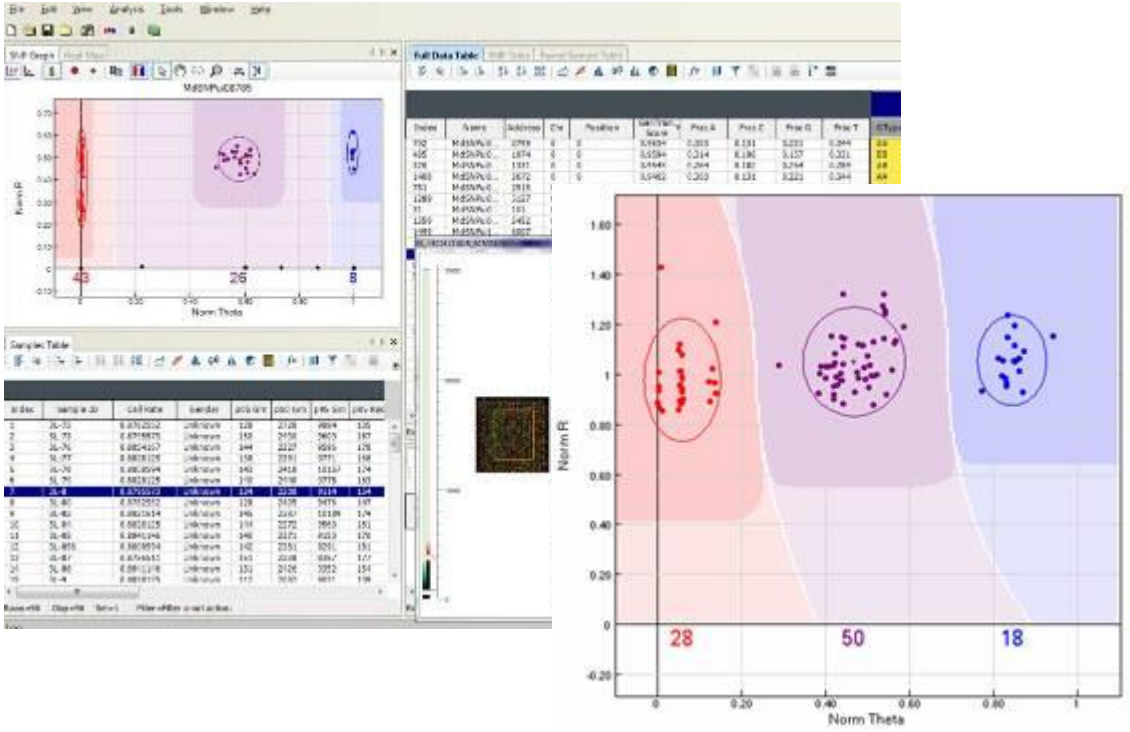
- SNP representation in a double stranded DNA fragment for 3 hypothetical individuals, tree 1, 2 and 3
- Tree 1 is heterozygous, whereas Trees 2 and 3 are homozygous (CTGN Short Course, University of California, 2009)

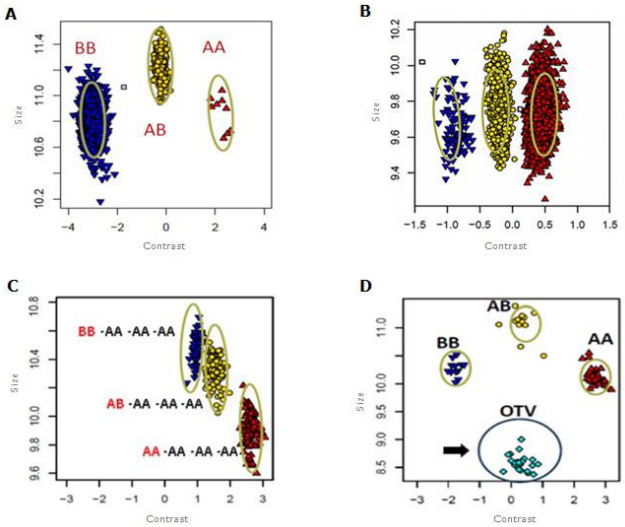


Call rate

국립보건연구원
National Institute of Health

- Genotype assignment
- The overall call rate of a sample is equal to the number of SNPs receiving an AA, AB, or BB genotype call divided by the total number of SNPs on the chip.





Genetic variance VS. Disease risk

국립보건연구원
National Institute of Health

genome.gov
National Human Genome Research Institute
National Institutes of Health

Google™ Search SEARCH

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

Home › Research Funding › Research Funding Divisions › Division of Genomic Medicine › GWAS Catalog

Share Print

Division of Genomic Medicine

A Catalog of Published Genome-Wide Association Studies

Division Staff | Funding Opportunities | Genomic Medicine Activities | GWAS Catalog | Meetings & Workshops | Potential Sample Collections for Sequencing | Programs | Publications | Trans-NIH Sequencing Inventory

Current uses of and future directions for the Genome-Wide Association Studies Catalog new

On Thursday, July 18th, 2013, the Division of Genomic Medicine held a webinar to highlight current uses and explore
presentations.

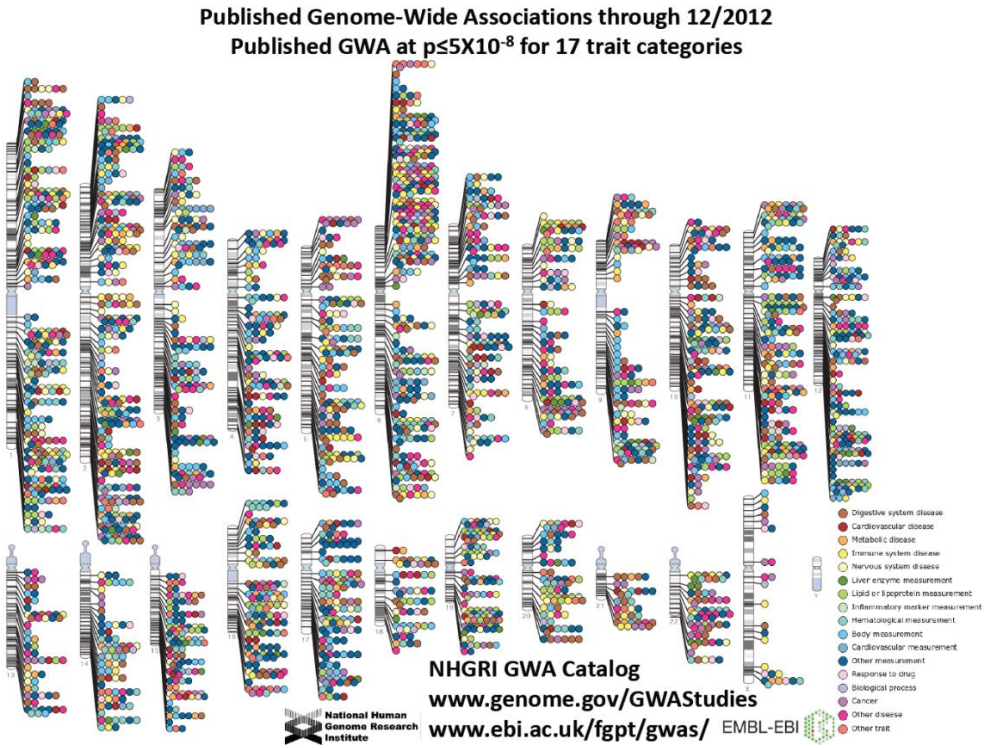
Additional information has been added to the HTML catalog columns below. For a description of column headin

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits
Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

View the Interactive Diagram new View the Full Catalog Download the Catalog Search

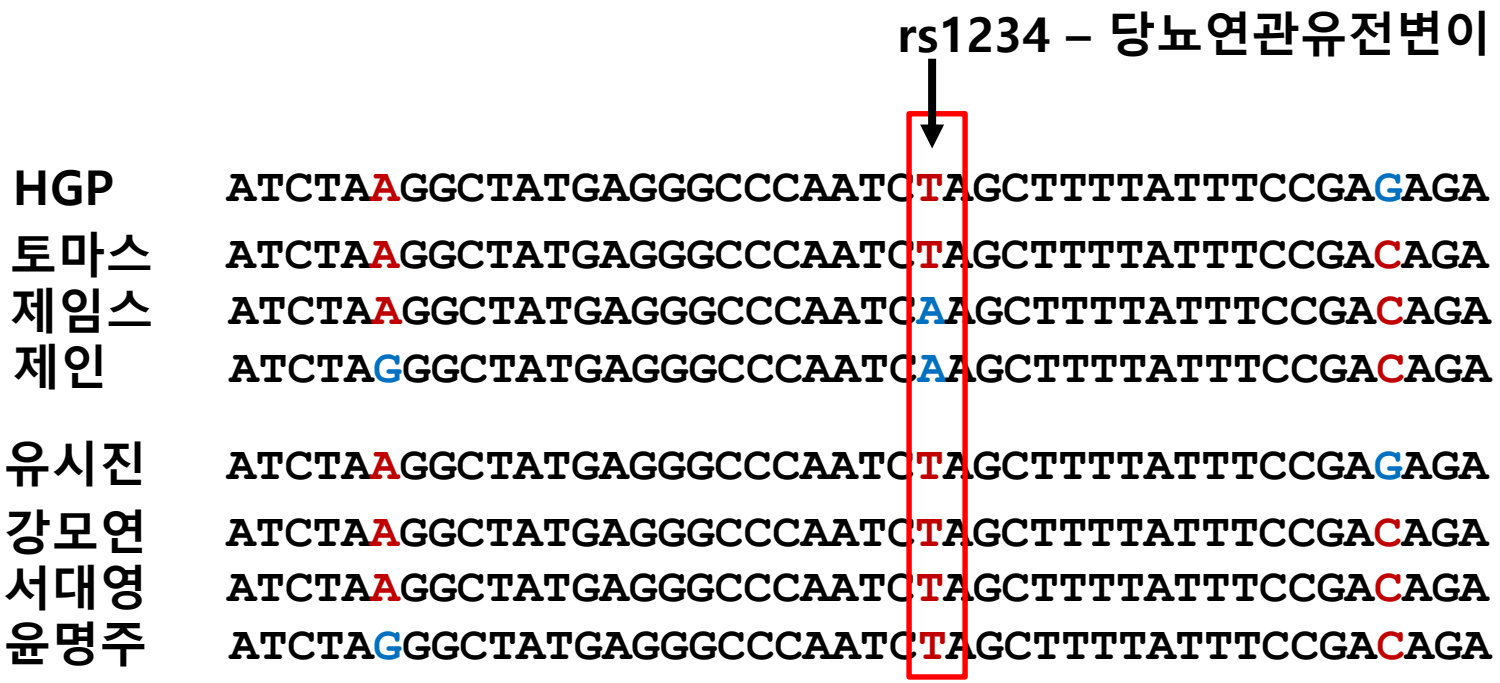
The genome-wide association study (GWAS) publications listed he
polymorphisms (SNPs) in the initial stage. Publications are organized fr
available. Studies focusing only on candidate genes are excluded from
daily NIH-distributed compilations of news and media reports, and occa
Navigator).

SNP-trait associations listed here are limited to those with p-values <
in p-values are rounded to the nearest single digit; odds ratios and alle
95 percent confidence intervals where applicable. Allele frequencies, p-
combined analysis (initial plus replication studies), are recorded below
For quantitative traits, information on % variance explained, SD increm
original paper are converted to OR > 1 for the alternate allele. Where



한국인유전정보기반 platform 필요

국립보건연구원
National Institute of Health



자국민 정보의 자원화

국립보건연구원
National Institute of Health

RARE GENETIC VA

UK 10K

biobank[uk]

Call Us On: 0800 0 276 276
Your feedback is important to us, tell us what you think

About | Participants | Resources | Approved research | Register & apply | Data Showcase | Scientists

You Are In | About UK Biobank

Int. J. Epidemiol. Advance Access published December 11, 2013

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2013;1–10
doi:10.1093/ije/dyt220

Parental diabetes and birthweight in 236 030 individuals in the UK Biobank Study

Jessica S Tyrrell,[1,2] Hanieh Yaghootkar,[2] Rachel M Freathy,[2] Andrew T Hattersley[3] and Timothy M Frayling[2]*

[1]European Centre for Environment and Human Health, University of Exeter Medical School, Truro, UK, [2]Genetics of Complex Traits, University of Exeter Medical School, Exeter, UK and [3]Molecular Genetics, Wonford Building, University of Exeter Medical School, Exeter, UK

*Corresponding author. Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK. E-mail: T.M.Frayling@exeter.ac.uk

ht, with the aim of improving the
esses – including cancer, heart
ns of dementia. UK Biobank
untry to take part in this project.
analysis, detailed information
build into a powerful resource to
not.

BBC NEWS HEALTH

Home | UK | Africa | Asia | Europe | Latin A

9 December 2013 Last updated at 01:33 GM

Hundred thousand
in Saudi Arabia

By Helen Briggs
BBC News

Saudi Arabia is launching a national research project to study the genetic basis of disease in its population

Up to 100,000 people in Saudi Arabia are to have their genetic codes mapped in a new human genome project.

Related Stories

View Cart

Tools

Follow u

Contents & Storage
Automation
Requirements & Compatibility
Downloads
Documentation & Literature

The HumanOmniZhongHua-8 BeadChip Kit uses the Infinium HD Super Assay and is compatible with the iScan and HiScan systems.

Complex disease study

국립보건연구원
National Institute of Health

- Caused by multiple genetic and environmental factors
- Ultimate goal of GWAS is to define genetic architecture of complex traits and disease and also provide new insight into disease pathophysiology (Genome Research. 2015, 25(10): 1432)

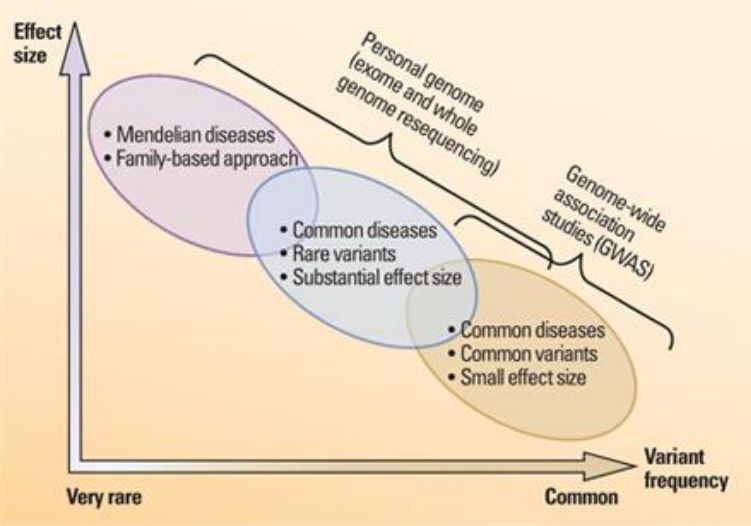
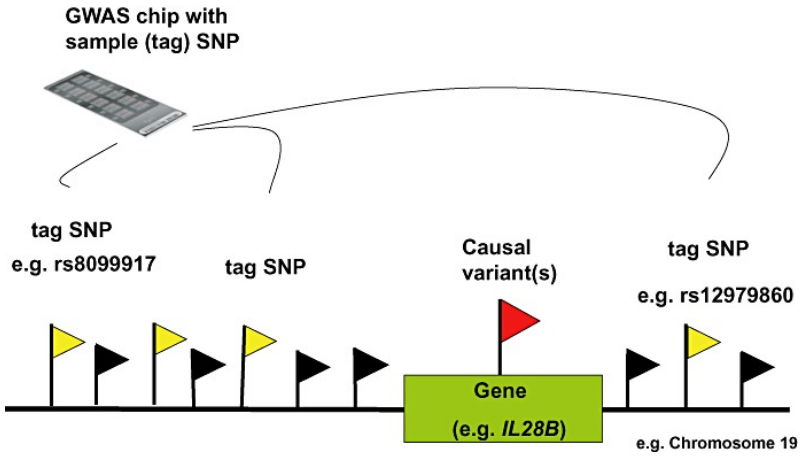
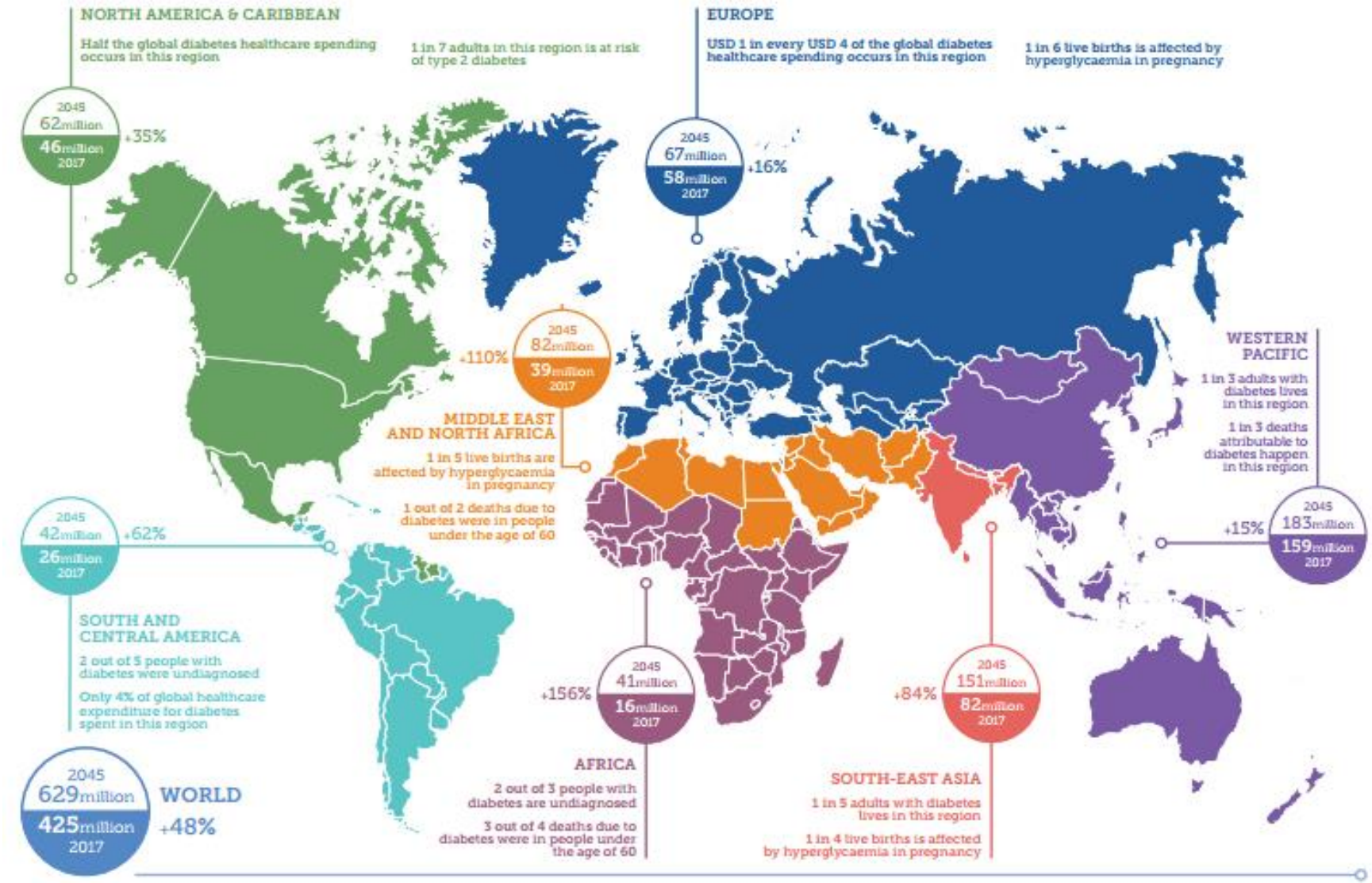


Image courtesy from http://www.sciencemag.org

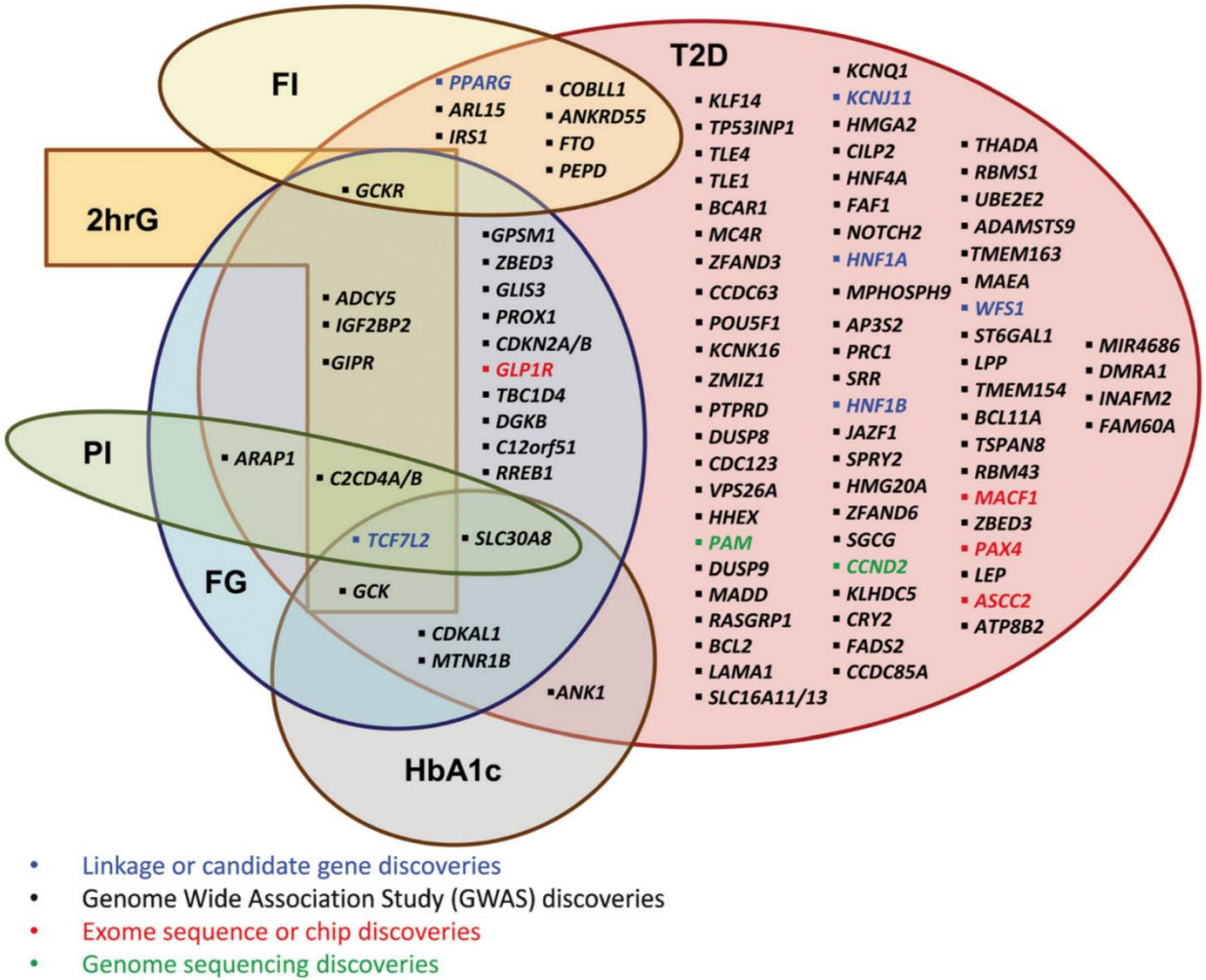


Journal of Gastroenterology and Hepatology (2012) 27(2):212-22



International Diabetes Federation Diabetes Atlas, 8th edition, 2017

Fig. 1 The Venn diagram contains loci significantly ($P < 5 \times 10^{-8}$) associated with type 2 diabetes (T2D) published before September 2016. The overlaps display significant ($P < 5 \times 10^{-8}$) overlapping associations reported in genome-wide association studies for each variant with other glycemic traits. The gene name provided is a label for the genetic locus and not meant to represent a causal gene. *FI* Fasting insulin, *2hrG* 2-hour glucose following an oral glucose tolerance test, *PI* Fasting proinsulin, *FG* Fasting glucose, *HbA1c* Hemoglobin A1c



Merino et al. Curr Diab Rep 2017

The power is all of numbers

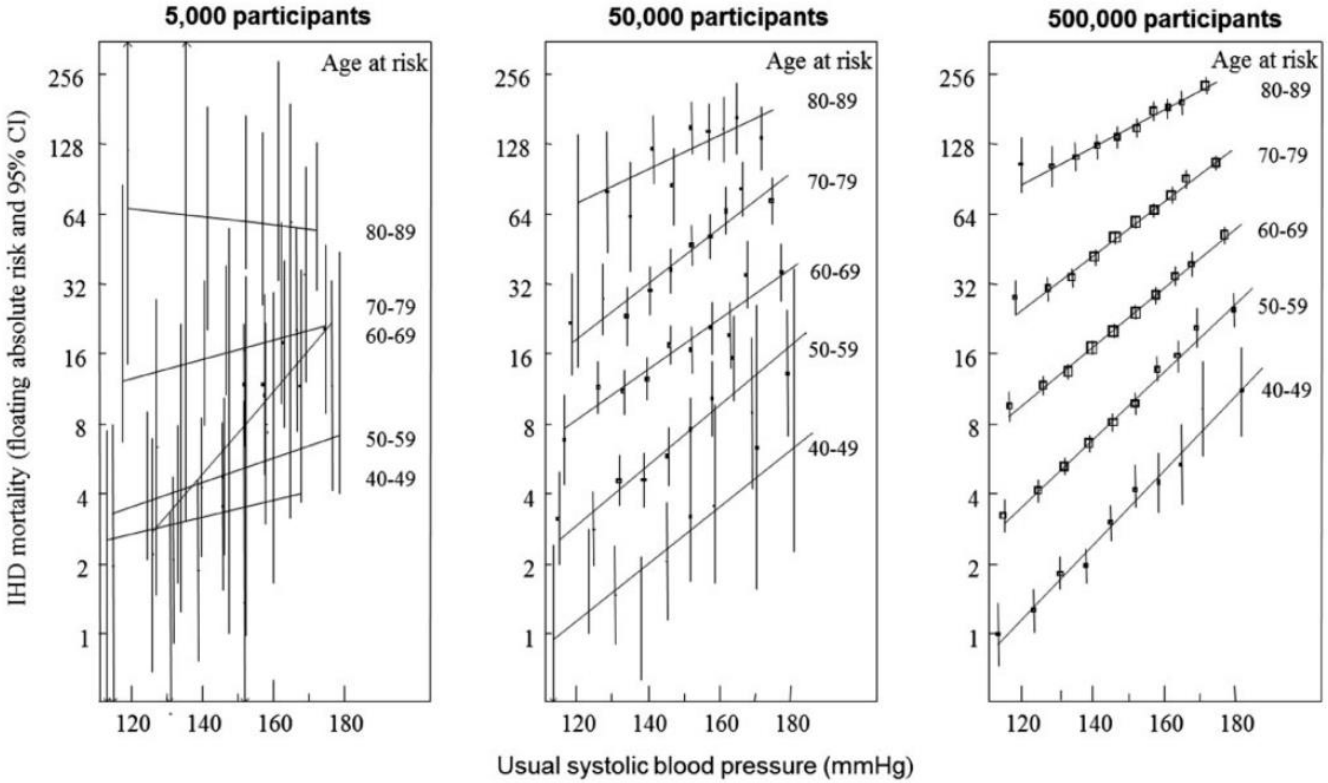


Figure 1 Absolute risk of ischaemic heart disease mortality by usual systolic blood pressure and age at risk in 5000, 50 000, and 500 000 participants. Unpublished figure containing data from the Prospective Studies Collaboration, obtained through personal communication. CI, confidence interval; IHD, ischaemic heart disease.

Eur Heart J. 2017. doi: 10.1093/eurheartj/ehx254

국가별 자국민 유전체칩 현황

국립보건연구원
National Institute of Health

국가	연도(시작)	샘플 수	사업명
미국	2011년	10만 명	UCSF-Kaiser RPGEH study
대만	2012년	5만 명	Taiwan Biobank Academia Sinica
영국	2013년	50만 명	UK Biobank
미국, 유럽 등	2013년	10만 명	iGeneTRAiN
미국	2013년	20만 명	Million Veteran Program

* UCSF: University of California San Francisco
* RPGEH: The Research Program on Genes, Environment, and Health
* iGeneTRAiN: The International Genetics & Translational Research in Transplantation Network

출처: Affymetrix

차세대 유전체칩

국립보건연구원
National Institute of Health

Arrays	Target Diseases	Main purpose	# of Contents	GW Tagging	Description
Metabo Chip	Metabolic Diseases	Replication Fine mapping	200K	No	-
Immuno Chip	Immune Diseases	Replication Fine mapping	200K	No	-
Exome Chip	Complex Diseases (Functional variants included)	Discovery	250K	No	-
Oncoarray	Cancers (5 cancers*)	Discovery Replication Fine mapping	530K	Yes	OncoArray Consortium 425,000 samples
UK BioBank	Complex Diseases	Discovery	820K	Yes	UK BioBank 500,000 samples
Kaiser BioBank	Complex Diseases	Discovery	650K	Yes	Kaiser BioBank 100,000 samples

***5 cancers: Breast, Ovarian, Intestine, Lung, Prostate**

한국인칩 == 만능열쇠???

ANALYSIS

nature genetics

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

Jian Yang[1,2,24]
Matthew R Rc
The LifeLines
Patrik K E Ma
Naomi R Wray

We propose a m
for human comp
genome sequen
based on whole-
~68% of variati
can be captured
method, we esti
all ~17 million i
(s.e.) = 2.3%) of
variance for bod
height- and BMI
selection. Consi
potential overes
based studies, h
and 30–40% for
small for both tr

With advances in genome sequencing technologies, it is now possible to sequence a human genome at high depth for $1,000, which, however, is still much more expensive than using a SNP array (for example, the Illumina CoreExome array). Given a fixed budget for genotyping and assuming that the genotyping cost using SNP arrays (for example, $50 per sample) is 20 times less than that for whole-genome sequencing (for example, $1,000 per sample), on average, 1000 Genomes Project imputation is currently at least 13 times more powerful than whole-genome sequencing using a multivariant association analysis approach (**Supplementary Fig. 15**). For a single variant–based association analysis, 1000 Genomes Project imputation is still at least 13 and 4 times more powerful than whole-genome sequencing in detect-

suggest that SNP array–based genotyping followed by imputation is now and in the near future will continue to be a more cost-effective strategy than whole-genome sequencing for GWAS of complex traits and diseases, even for rare variant associations. Nevertheless,

the analyses above compared the average power for variants in a certain MAF range. There are a number of sequence variants (~10% rare and ~1% common) that are almost not tagged by any imputed variant individually (single-variant tagging $r^2_{max} < 0.05$) in 1000 Genomes Project imputation based on the Illumina CoreExome array (**Supplementary Fig. 16**). For association analysis of such variants and those with extremely low frequency or unique to specific populations, high-coverage whole-genome sequencing or a haplotype-based method will be a more efficient strategy. In contrast, it has been suggested that extremely low-coverage whole-genome sequencing followed by imputation can be even more cost-effective than SNP array–based imputation for common variants[35], an interesting strategy that is worth being further investigated for its performance on rare variants.

With the latest imputation reference panel of large sample size (n = 31,000; Haplotype Reference Consortium, personal communication) and very large GWAS cohorts genotyped on the same type of

for example, the UK Biobank has genotyped >400,000
sing Affymetrix Axiom arrays) that are soon becoming
can expect a great improvement in imputation accuracy
nts. For complex traits and diseases that have a genetic
imilar to that of height (enrichment of height-associated
MAF <0.1), we can expect to see a wave of discovery
of trait- or disease-associated low-MAF variants in the near future, without the need for large-scale whole-genome sequencing.

한국인칩 제작

한국인칩 제작배경 : 기존 연구기법의 한계

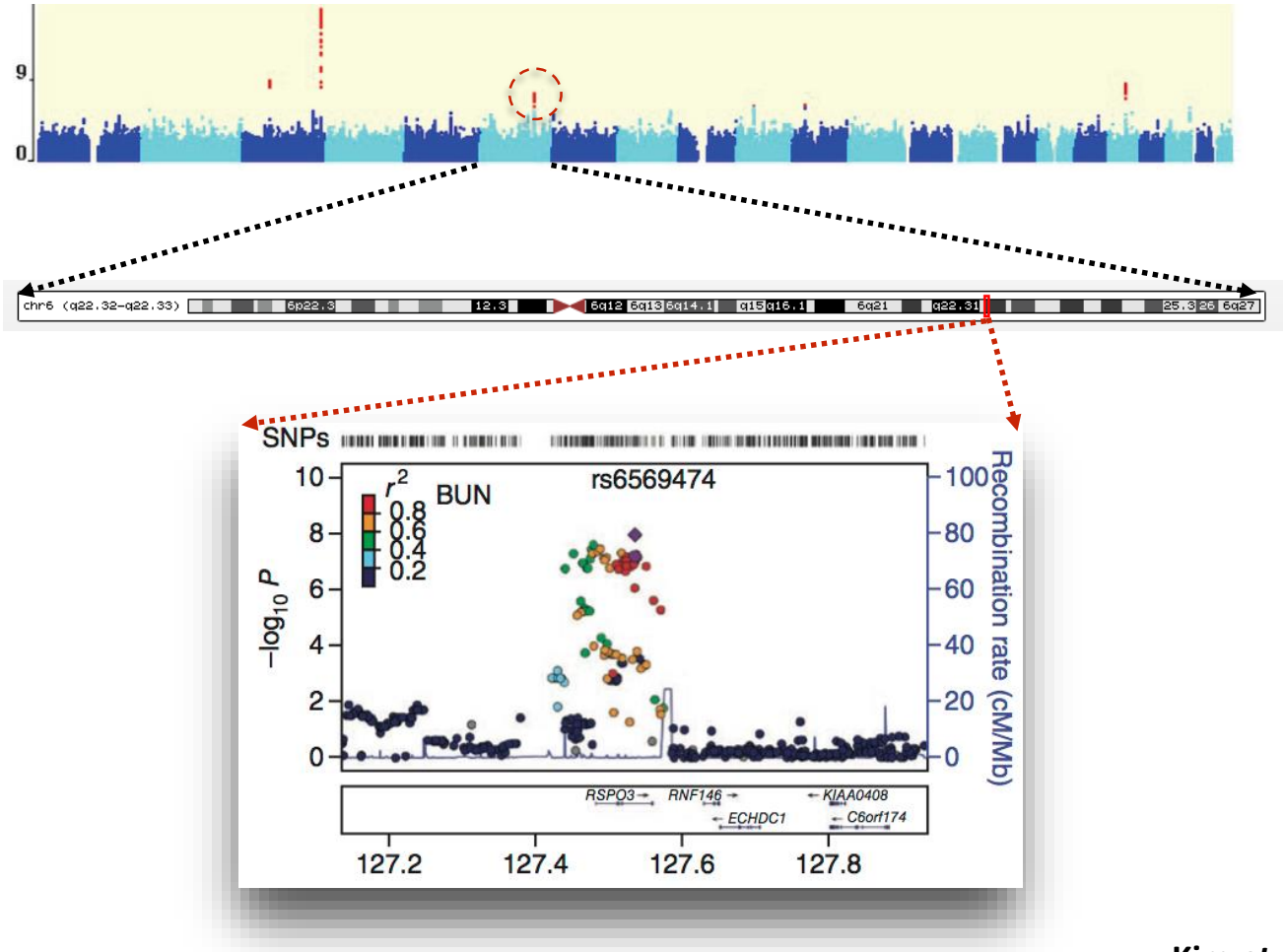
국립보건연구원 National Institute of Health

- 한국인 만성질환 유전체 연구를 위한 대규모 인구집단 유전체 연구의 기존 연구기법의 문제점 대두
  - 유전변이 칩: 서양인 중심 설계, 한국인 염기서열정보 미반
    - *낮은 Genomic coverage (~75%, 1KG ASN, MAF 5% 기준)
  - 차세대염기서열분석 기법
    - * 높은 계산력과 유전변이 칩 대비 수십 배의 분석 시간 요구

- 이러한 한계 극복을 위한 인종 특이칩 제작
  - 인종 별 염기서열 정보 기반, 각 인종의 질환 유전체 연구에 최적화
    - * 인종별 1000게놈 프로젝트 phase 3 서양인(503명), 동아시아인(504명)
  - 낮은 비용 (기존칩 대비 약 3-5배, NGS 대비 약 10배 절감)

Example of Genome-wide scan

- High genomic coverage confers high association mapping power



Kim *et al.* Nature 2011

ARTICLE

Singapore Sequencing Malay Project

Deep Whole-Genome Sequencing of 100 Southeast Asian Malays

Lai-Ping Wong,[1,14] Rick Twee-Hee Ong,[1,14] Wan-Ting Poh,[1,14] Xuanyao Liu,[1,2,14] Peng Chen,[1] Ruoying Li,[1] Kevin Koi-Yau Lam,[1] Nisha Esakimuthu Pillai,[3] Kar-Seng Sim,[4] Haiyan Xu,[1] Ngak-Leng Sim,[4] Shu-Mei Teo,[1,2] Jia-Nee Foo,[4] Linda Wei-Lin Tan,[1] Yenly Lim,[1] Seok-Hwee Koo,[5] Linda Seo-Hwee Gan,[6] Ching-Yu Cheng,[1,10,11] Sharon Wee,[1] Eric Peng-Huat Yap,[6] Pauline Crystal Ng,[4] Wei-Yen Lim,[1] Richie Soong,[7] Markus Rene Wenk,[8,9] Tin Aung,[10,11] Tien-Yin Wong,[10,11] Chiea-Chuen Khor,[1,4,10,12] Peter Little,[3] Kee-Seng Chia,[1] and Yik-Ying Teo[1,2,3,4,13,*]

- Variant discovery
- LOF variants
- Population Structure
- Mutation hotspot
- Impact of Sequencing Coverage
- Accessing Genomic coverage of microarray
- Comparison of Reference Panels in Genotype imputation

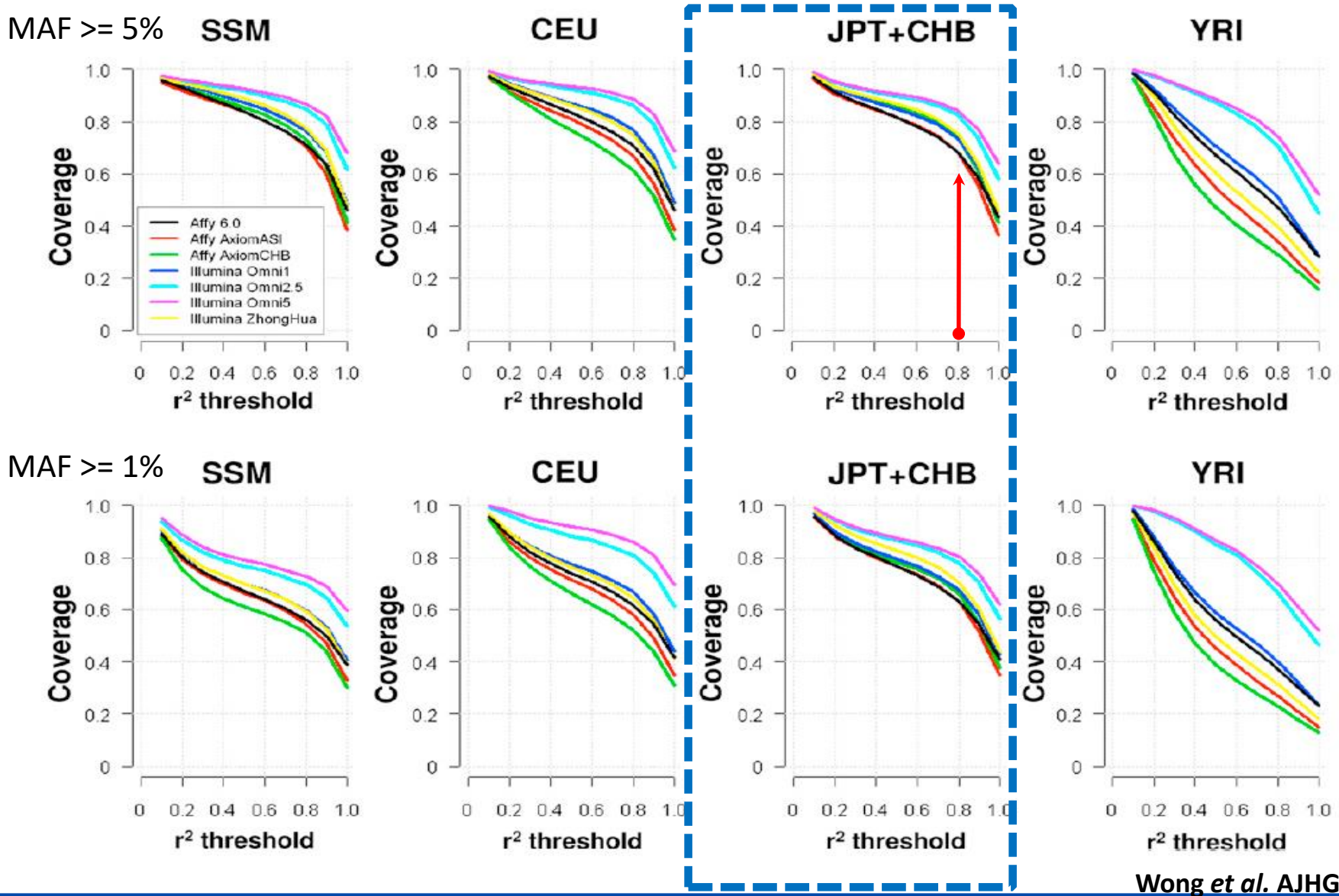
$$\text{Genomic Coverage} = \frac{\text{\# of Tagged markers}}{\text{Total \# of SNP}}$$

Wong *et al.* AJHG 2013

기존 칩 제작 전략

- Evenly spaced markers
  - Affymetrix 500K, 5.0
- Tagging SNP markers
  - Illumina SNP chips
- Hybrid approach (Evenly spaced + Tagging SNP)
  - Affymetrix 6.0

Hao et al. PLoS Genet 2008
http://www.affymetrix.com
http://www.illumina.com



Wong *et al.* AJHG 2013

한국인칩 컨텐츠 제작 목표

(1) **Minimize the number of markers filtered by QC** because of ethnic difference, resulting in maximum utilization of KoreanChip;

(2) Include the highest possible amount of **potentially damaging variants observed in Koreans** that can directly affect coding sequence;

(3) Achieve **higher imputation-based genomic coverage** at common and rare variants;

(4) Ensure **cost-effectiveness** to provide more genomic information on the same budget to facilitate genome–phenome studies.

한국인칩 성능평가

국립보건연구원
National Institute of Health

KoreanChip (833K)

Evaluation

Reproducibility	*Accuracy*	*Chip Contents comparison*	*Genomic coverage*	*Utility (GWAS)*
• Genotype comparison between 35 blind duplicates of the KoreanChip from different batches	• Genotype concordance test of identical genotype between the KoreanChip and previously reported data	• Contents comparison between the KoreanChip and existing commercial arrays - Number of shared markers - Number of functional markers on each platforms	• Calcaulation of genomic coverage between the KoreanChip and existing arrays - comparison with well-known commercial arrays - comparison with next-generation arrays using same individuals (n=96)	• Preliminary GWAS of blood biochemical traits using the KoreanChip (n=6,949) - HDL, LDL, TG, ALT, AST • Follow-up replication analysis using directly genotyped for significant variants with TaqMan-based assay (n=6,000)

Ansan and Asung study
- KoreanChip (6,949)
- KoreanChip (96) randomly selected from 6,945
- AFFY 5.0 (6,949)
- ILMN Exome array (5,793)
- Exome sequencing (155)

HEXA study
- AFFY 6.0 (3,695)
- TaqMan qPCR (6,000)
- Axiom Biobank array (96)
- Axiom UKB (96)
- Axiom PMRA (96)
- ILMN GSA (96)

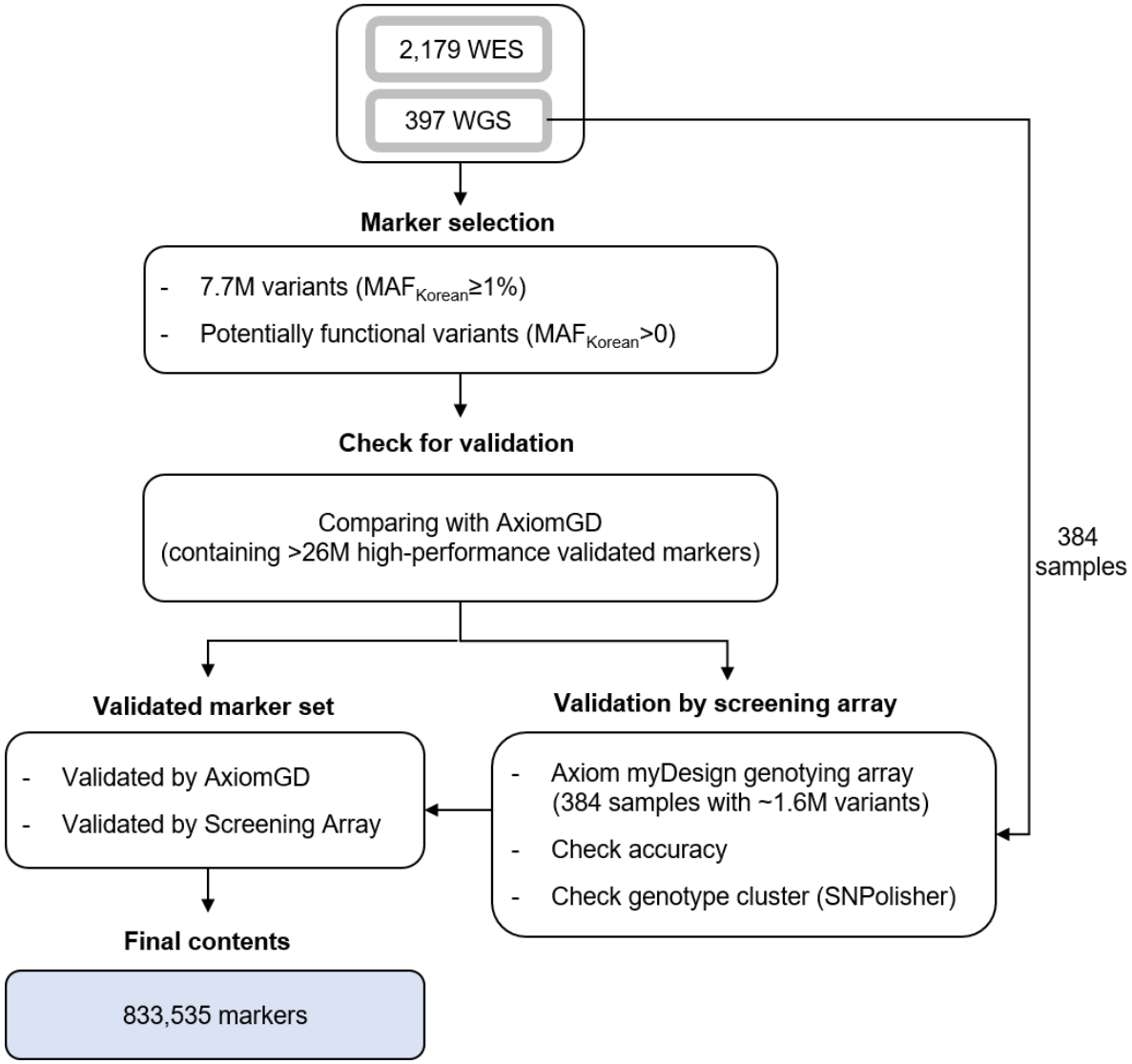
CAVAS study
- ILMN Omin 1 (3,666)

Ansan and Asung study
- KoreanChip (6,949)
- KoreanChip (96) randomly selected from 6,945
- AFFY 5.0 (6,949)
- ILMN Exome array (5,793)
- Exome sequencing (155)

HEXA study
- AFFY 6.0 (3,695)
- TaqMan qPCR (6,000)
- Axiom Biobank array (96)
- Axiom UKB (96)
- Axiom PMRA (96)
- ILMN GSA (96)

CAVAS study
- ILMN Omin 1 (3,666)

마커 선별 방법 (2014. 5 ~ 8)



정확도

Methods-Accuracy and reproducibility

국립보건연구원
National Institute of Health

S4 Table. Comparison of accuracy between KCHIP and other platforms

Platform	Overlapping with KCHIP, N		Accuracy, %	
	Subject	Marker	Overall	Hetero
Affymetrix Genome-wide human SNP array 5.0	6,949	41,246	99.8	99.5
Illumina HumanExome BeadChip v1.1	5,793	34,683	99.9	99.7
Exome sequencing (Illumina Hiseq 2000)	155	90,020	99.8	99.7

Accuracy: # of True genotypes / # of Total genotypes

Overall: Overall accuracy, Hetero: Accuracy of heterozygotes

Reproducibility (duplicate blind comparisons, 35 samples in different batches): 99.77%.

컨텐츠

KoreanChip content– Functional Variant

국립보건연구원
National Institute of Health

Table 1. Contents summary of KoreanChip

Category	Number of SNPs*	Contents (%)
Tag SNPs for genome-wide coverage	600,294	72.02
Functional loci (nonsynonymous SNPs and Indels)	208,039	24.96
eQTL	16,690	2.00
HLA	6,659	0.80
Fingerprint	255	0.03
NHGRI GWAS catalog	7,811	0.94
KIR	1,544	0.19
Pharmacogenetics/ADME	1,881	0.23
Common mitochondrial DNA variants	178	0.02
Y chromosome markers	806	0.10
Total	833,535	-

*Some SNPs are overlapped among categories.

eQTL, expression Quantitative Trait Loci; HLA, Human leukocyte antigen; KIR, Killer cell immunoglobulin like receptors; ADME, Absorption, Distribution, Metabolism, and Excretion.

사용칩과의 컨텐츠 비교

국립보건연구원
National Institute of Health

Chip Contents comparison

- Contents comparison between the KoreanChip and existing commercial arrays
  - Number of shared markers
  - Number of functional markers on each platforms

S5 Table. Contents comparison with existing arrays

Platform	AFFY5.0	AFFY6.0	ILLU 1M
KoreanChip	47,846	90,057	123,761
AFFY5.0	-	482,398	140,046
AFFY6.0	-	-	271,989
ILLU 1M	-	-	-

S6 Table. Contents comparison with next-generation arrays

Platform	Axiom Biobank	UK Biobank	ILMN Exome	PMRA
KoreanChip	219,690	238,929	42,807	275,312
Axiom Biobank	-	398,587	229,317	244,305
UK Biobank	-	-	82,225	286,215
ILMN Exome	-	-	-	34,348
PMRA	-	-	-	-

기능변이

국립보건연구원
National Institute of Health

Table 2. Comparison of contents between KoreanChip and other genotyping chips

Platform	Total marker N	Annotated marker[1)] N	Nonsyn marker[2)] N (%)	ASN marker[3)] N (%)
Affymetrix 5.0	500,568	489,457	2,179 (0.4)	769 (0.2)
Affymetrix 6.0	934,969	892,584	4,889 (0.5)	1,750 (0.2)
Illumina 1M	1,099,726	1,066,324	45,832 (4.3)	12,516 (1.2)
Illumina Exome array	242,761	241,923	217,775 (90.0)	39,480 (16.3)
Illumina GSA	700,078	688,062	87,759 (12.8)	21,371 (3.1)
Axiom Biobank	718,212	645,060	251,080 (38.9)	46,416 (7.2)
Axiom UK Biobank	845,487	823,336	104,058 (12.6)	19,487 (2.4)
Axiom PMRA	920,744	856,797	44,819 (5.2)	6,088 (0.7)
KoreanChip	**833,536**	**829,635**	**183,607 (22.1)**	**89,413 (10.8)**

1) annotated by snpEff v4.1d based on the database of dbNSFP2.7 (functional prediction and annotation of nonsynonymous marker

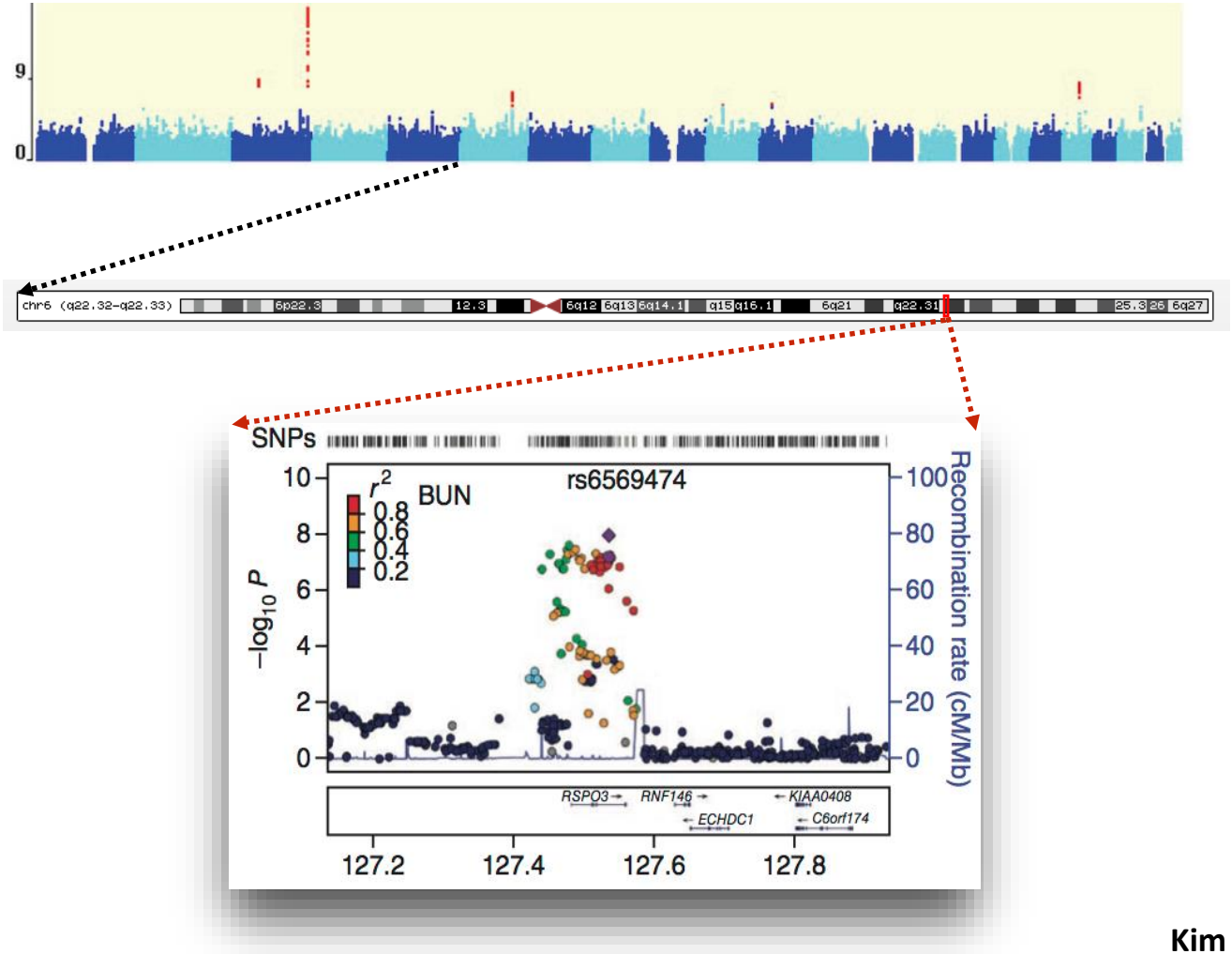
2) proportion of nonsynymous markers among annotated markers

3) proportion of nonsynonymous makers, damaging ≥1, and allele frequency > 0 observed in East Asian ancestry among annotated markers

Genomic coverage

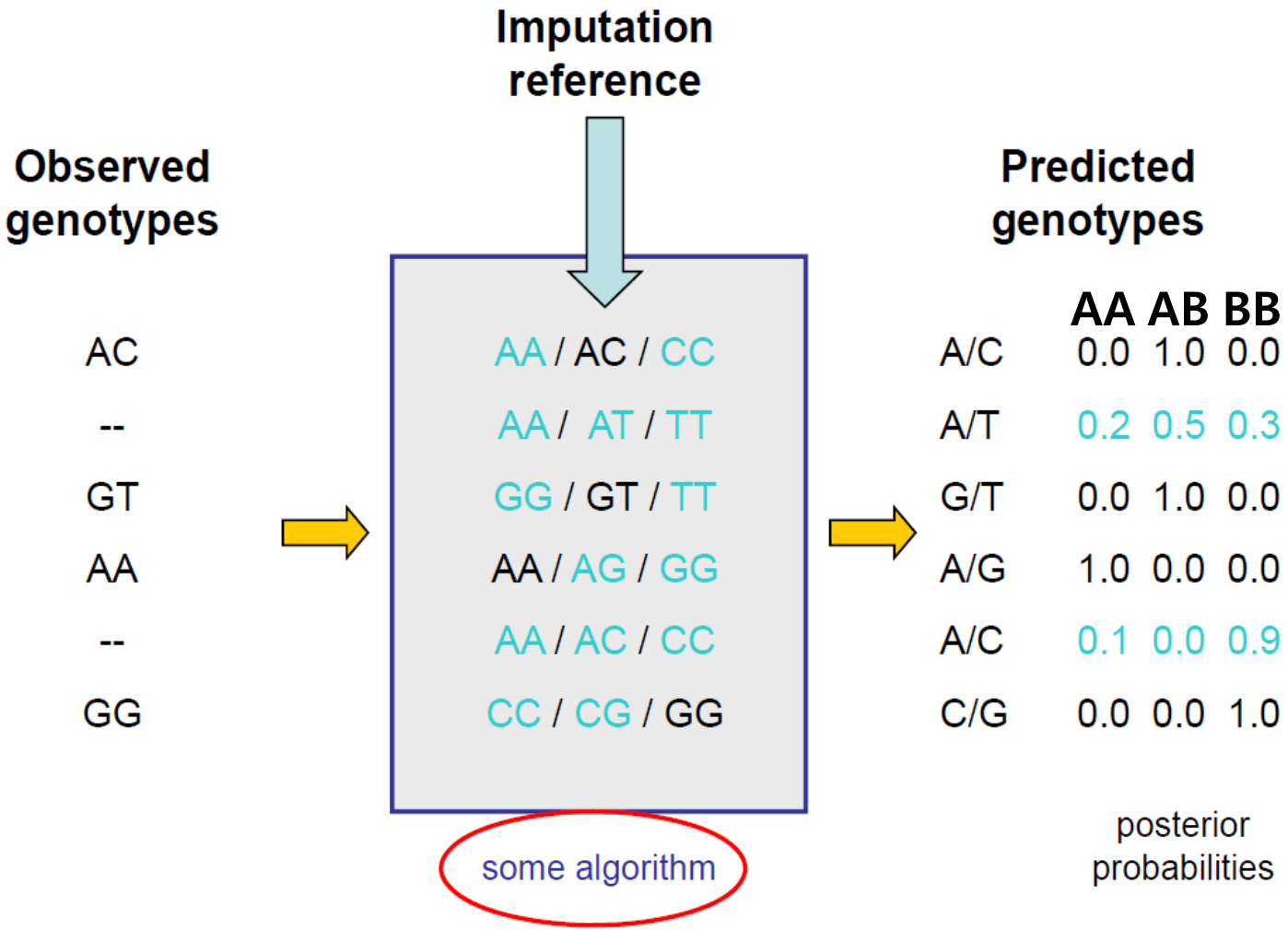
Example of Genome-wide scan

- High genomic coverage confers high association mapping power



Kim *et al.* Nature 2011

Introduction – Imputation



Imputation GWAS grid (UK Biobank)

3. Genome-wide Coverage

3.1 Genome-wide coverage for common variants (348,569 markers)

348,569 markers were selected using Affymetrix' imputation aware marker choice algorithms (Hoffman et al, Genomics 98 (2011) 422–430) to provide genome-wide coverage in Caucasian European populations of common (EMAF≥5%) markers (using the EUR panel defined as the GBR, CEU, FIN, IBS and TSI samples from 1000G). This explicitly included the set of 246,055 markers on Affymetrix' Axiom Biobank Genotyping Array selected to capture common (EMAF≥5%) variation.

3.2 Genome-wide coverage for low frequency variants (280,838 markers)

280,838 markers were selected using Affymetrix' imputation aware marker choice algorithms to provide genome-wide coverage in Caucasian European populations of low frequency (1%<EMAF<5%) markers (using the EUR panel described above).

Genome-wide imputation coverage in the EUR panel (see above for definition) estimated by Affymetrix:

Category	EMAF range	Mean r^2	% of markers with r^2>0.8
Common	5%≤EMAF≤50%	0.92	90.1%
Low frequency	1%<EMAF<5%	0.785	67.1%

Estimated genomic coverage

- Genomic Coverage
  - Genomic Coverage: the proportion of variants captured by a genotyping microarray (Nelson et al. G3 2013)
  - Imputation based genomic coverage: fraction of variants with imputation quality score ≥ 0.8
- Imputation
  - Reference panel: 1,000 genomes project phase 3 (2,504 samples)
  - Imputation: Impute v2.3

Platform	# of markers	# of samples
AFFY 5.0	500K	8,842
AFFY 6.0	900K	3,703
Illumina 1M	1M	3,667
KORV1.0	833K	7,000

Genomic coverage

국립보건연구원
National Institute of Health

Genomic coverage

- Calcaulation of genomic coverage between the KoreanChip and existing arrays
  - comparison with well-known commercial arrays
  - comparison with next-generation arrays using same individuals (n=96)

Table 3. Comparison of genomic coverage

Platform	Allele frequency			
	# of samples	MAF≥0.01	Common (MAF≥0.05)	Less common (0.01≤MAF<0.05)
KoreanChip	**6,949**	**89.86**	**95.38**	**73.65**
Affymetrix 5.0	6,949	76.25	84.78	51.23
Affymetrix 6.0	3,695	83.93	91.67	61.23
Illumina Omni 1M	3,666	86.97	94.10	66.01
KoreanChip	**96**	**88.37**	**95.24**	**68.22**
Axiom Biobank	96	81.94	91.56	53.74
UK Biobank	96	85.21	94.05	59.30
Axiom PMRA	96	87.09	94.48	65.42
Illumina GSA	96	84.38	92.27	61.24

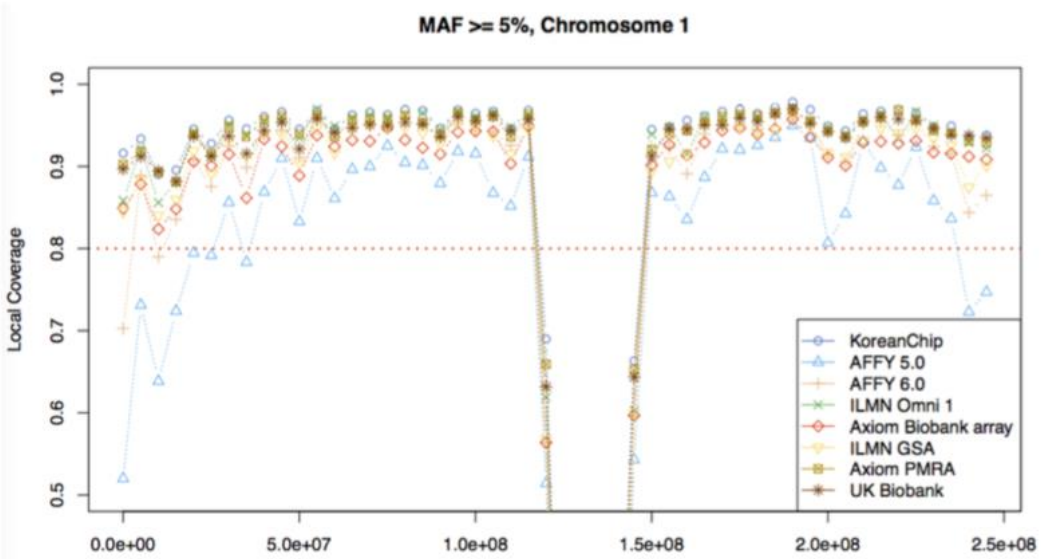
* Calculated using imputed data

** Representative chips of next-gen arrays: Axiom PMRA (Precision Medicine Research Array), UK Biobank, Illumina GSA (Global Screening Array), and Axiom Biobank

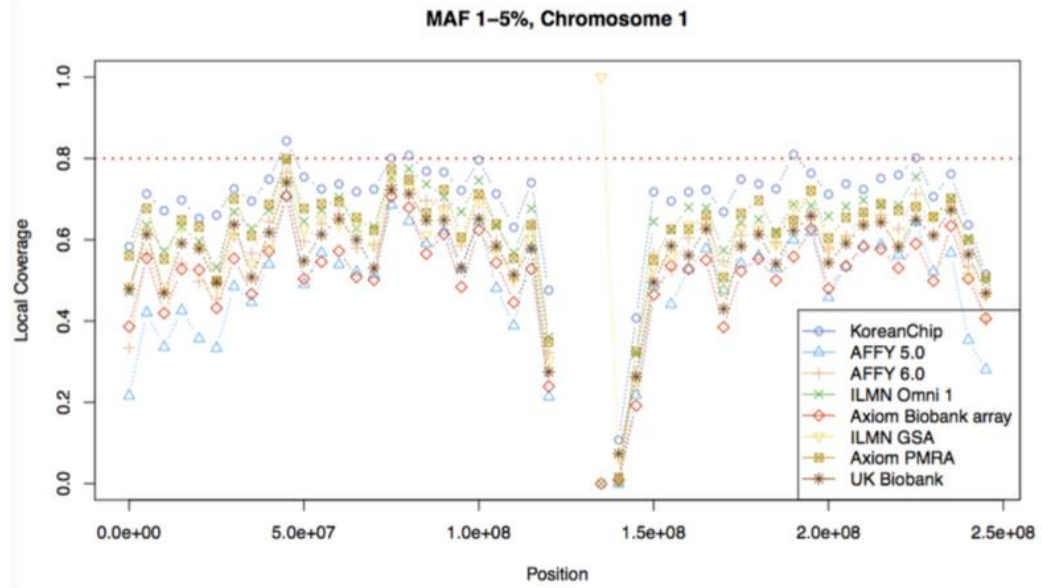
Estimated genomic coverage

국립보건연구원
National Institute of Health

CHR 1
(MAF ≥ 5%)



CHR 1
(MAF 1-5%)



연관성 분석 결과

UK BiLEVE: Aims

Phenotypes define by extremes of the lung function distribution

- heavy smokers
- non-smokers

Common SNPs
MAF>5%

Low frequency SNPs
MAF 1-5%

Rare SNPs
MAF <1%

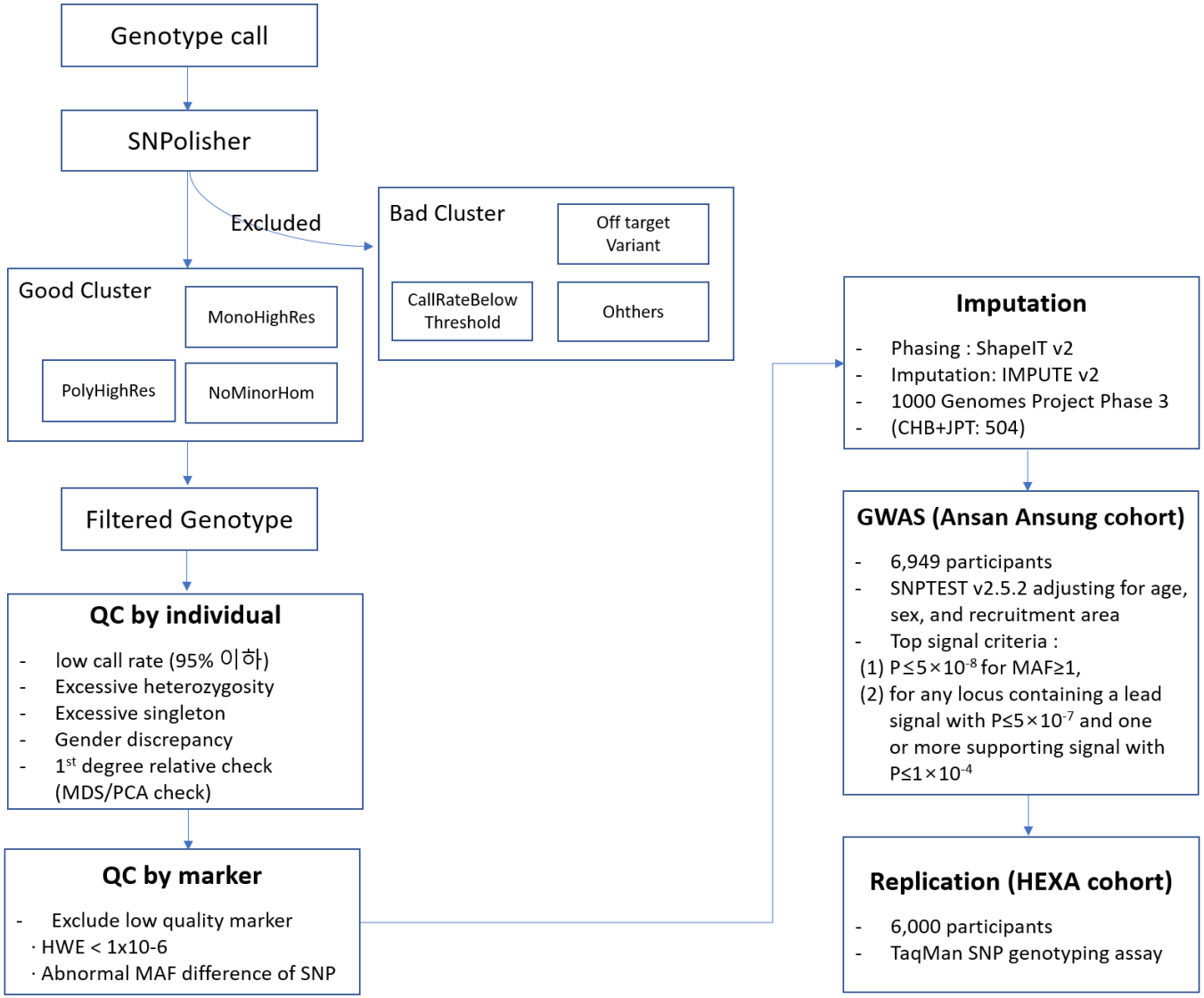
Putative functional variation

Novel regions

Fine map known regions

Independent signals in known regions

Overall scheme of GWAS



Comparison results of Association signals

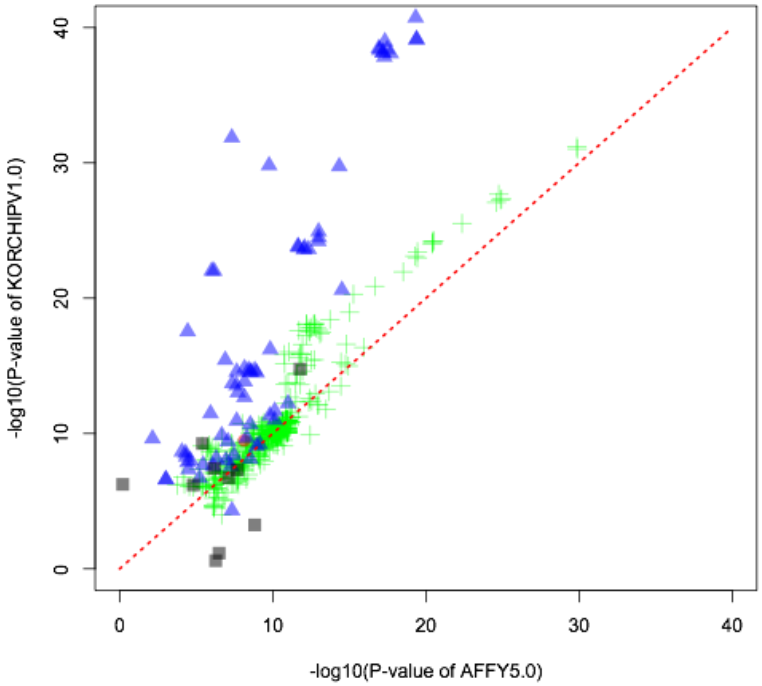
국립보건연구원
National Institute of Health

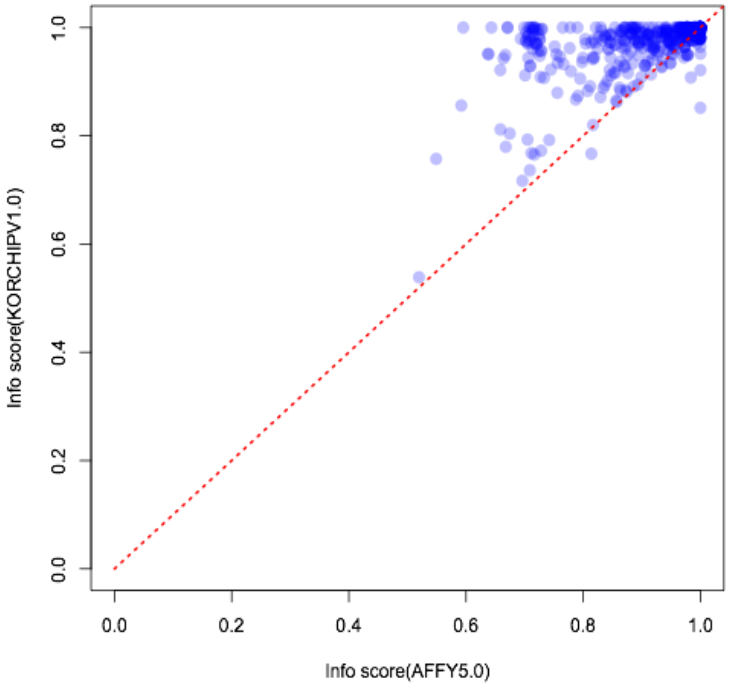
- Comparison analysis
  - Data: AFFY5.0, KORV1.0 identical 7,000 samples
    (Imputed using 1KG phase3, 8,700,150 variants)
  - Phenotype: Lipids (HDL, LDL, TG), Liver enzyme (AST, ALT, GGT), T2D
  - Association test: SNPTEST v2.5
  - Covariates: age, gender, recruitment area
  - Top signal selection
  - P-value ≤ 10-6 (Lipids)

Comparison results of Association signals

국립보건연구원 National Institute of Health

- Association results (HDL)
  - High quality (info score > 0.8): similar association results
  - In overall, K-CHIP showed higher imputation quality and stronger statistical significance

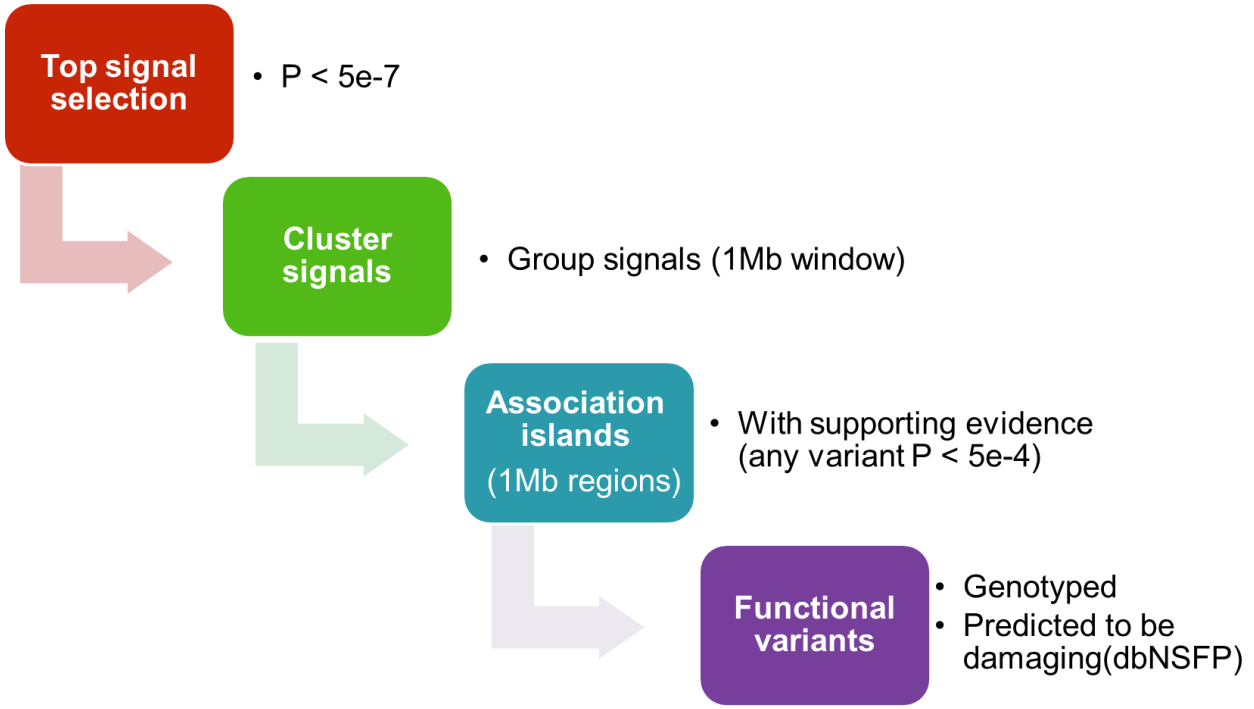




Preliminary association analysis

국립보건연구원 National Institute of Health

- Discovery: 7,000 samples KCHIP (Imputed using 1KG phase3)
- Replication: 6,000 samples (Taqman genotyping)
- Phenotype: Lipids (HDL, LDL, TG), Liver enzyme (AST, ALT, GGT)



31 variants remained

Application to GWAS (known or novel variants)

국립보건연구원 National Institute of Health

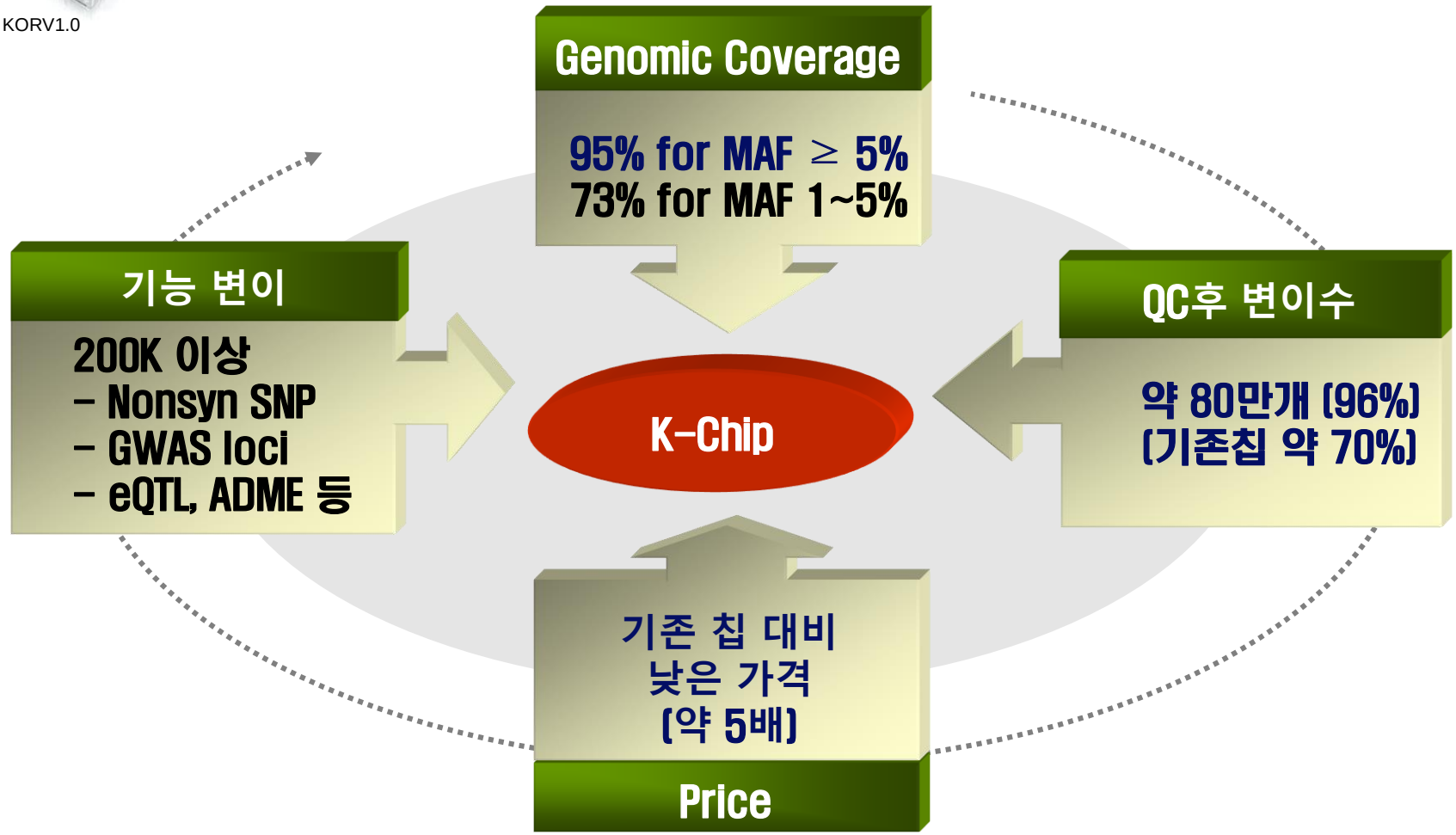
Gene	Trait(s)	EAF(%) KOR	gnomAD EAS	gnomAD EUR	gnomAD AFR	Discovery (~6,949 samples) Beta(SE)	Discovery (~6,949 samples) P-value	Replication (~6,000 samples) Beta(SE)	Replication (~6,000 samples) P-value
5 variants at known loci									
-	TG	33.31	37.00	20.86	21.53	-0.0415(0.0089)	3.27E-06	-0.0483(0.0105)	4.26E-06
C2orf16	TG	52.87	47.81	27.16	6.61	0.0379(0.0084)	7.20E-06	0.0560(0.0100)	2.36E-08
BUD13	HDL	6.61	7.22	6.06	1.16	0.0330(0.0073)	7.04E-06	0.0229(0.0081)	4.66E-03
C19orf80, DOCK6	LDL	27.31	25.93	4.42	18.05	-0.0203(0.0056)	3.16E-04	-0.0281(0.0058)	1.57E-06
	TCHL	27.02				-3.8231(0.6689)	1.14E-08	-3.6170(0.7294)	7.29E-07
APOE	LDL	37.47	39.62	63.57	85.81	-0.2010(0.0052)	1.23E-04	-0.0210(0.0055)	1.31E-04

ALT lowering variants (missense)
a reduction in ALT level of 7.0% (1.982 IU/L) and 5.9% (1.658 IU/L) of the mean value

APOB	LDL	0.97	0.26	0	0	0.1509(0.0259)	5.87E-09	0.1117(0.0256)	1.27E-05
	TCHL					15.9680(3.1140)	3.01E-07	13.2300(3.2040)	3.69E-05
2 novel associations of a known variant (Asian-specific)									
ALDH2	ALT	15.67	25.65	0.002	0.02	-0.0586(0.0107)	4.98E-08	-0.0481(0.0114)	2.86E-05
	AST					-0.0541(0.0075)	5.20E-13	-0.0372(0.0075)	8.14E-07
2 novel variants at novel loci (Asian-specific)									
GPT	ALT	0.12	0.10	0.004	0	-0.6843(0.1140)	2.02E-09	-0.5574(0.1023)	5.30E-08
GPT	ALT	0.14	0.11	0	0	-0.5058(0.1048)	1.41E-06	-0.4972(0.1024)	1.24E-06

한국인칩 장점

KORV1.0



요약

- KCHIP contains tagging SNPs and functional variants
  - Higher genomic coverage than commercial chips
  - Discovered functional variants in the previously reported regions
  - Discovered novel rare associations

- Customized chips can help to discover novel loci (Wain et al. 2015, UK BiLEVE)
  - not detected in previous because it was neither directly genotyped nor imputed with sufficient quality

- Association power will be maximized by various sampling from a large biobank