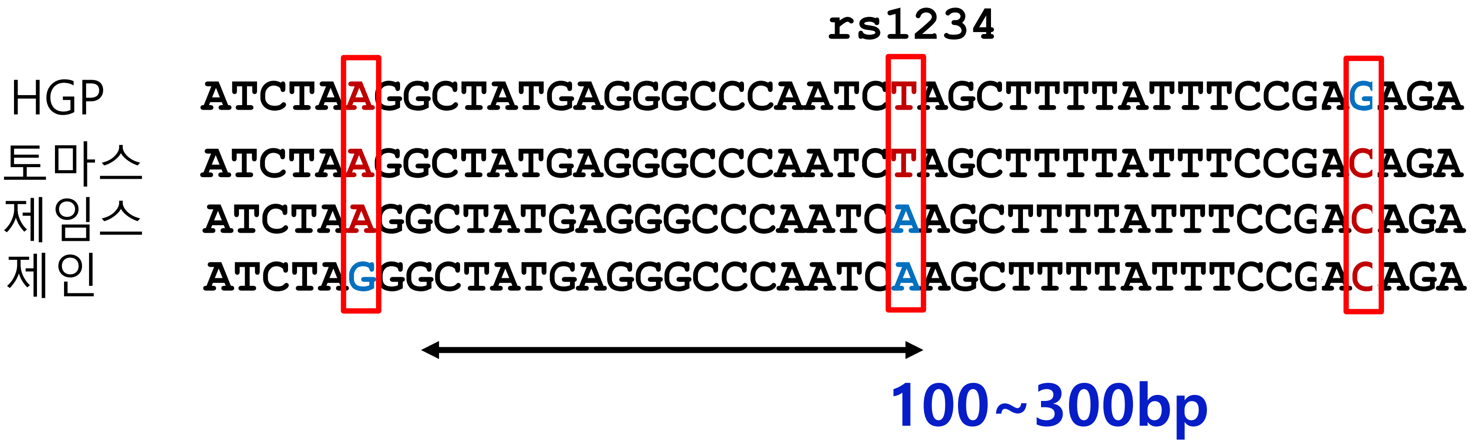
한국인칩 컨텐츠 특징 및 성능 소개

국립보건연구원 미래의료연구부 유전체연구기술개발과

2022 통계유전학워크샵: 한국인칩을 이용한 대규모 유전체정보 분석

Single Nucleotide Polymorphism (SNP)

- 대표적인 유전변이
- 평균 100~300bp 당 1개의 단일염기 차이 발생
- 인간이 가지고 있는 유전변이 중 **가장 많이 존재하는 형태** (전체변이의 90%)
- 한 사람의 인간은 2~3백만 개 이상의 SNP을 갖고 있는 것으로 알려짐
- 발굴된 SNP은 미국 NIH 산하 NCBI의 dbSNP 데이터베이스에 저장

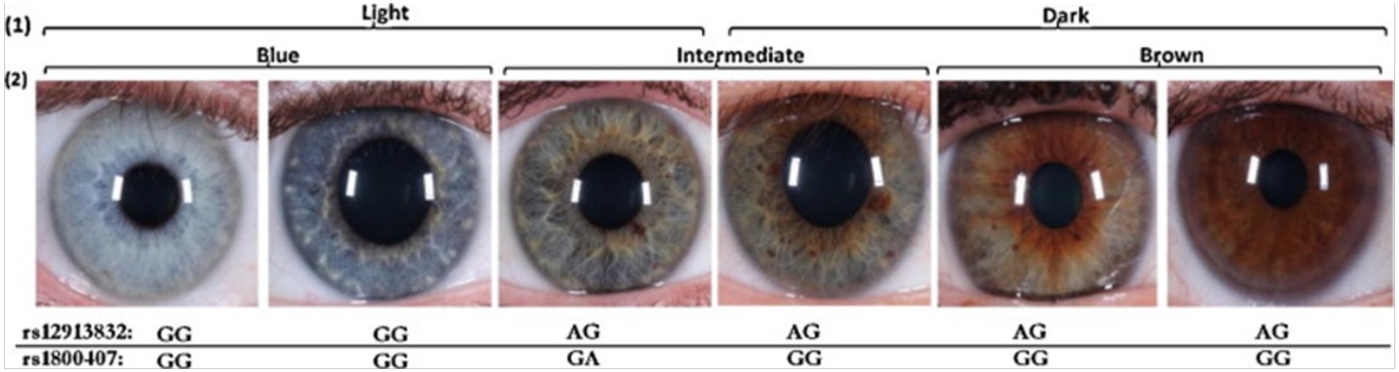


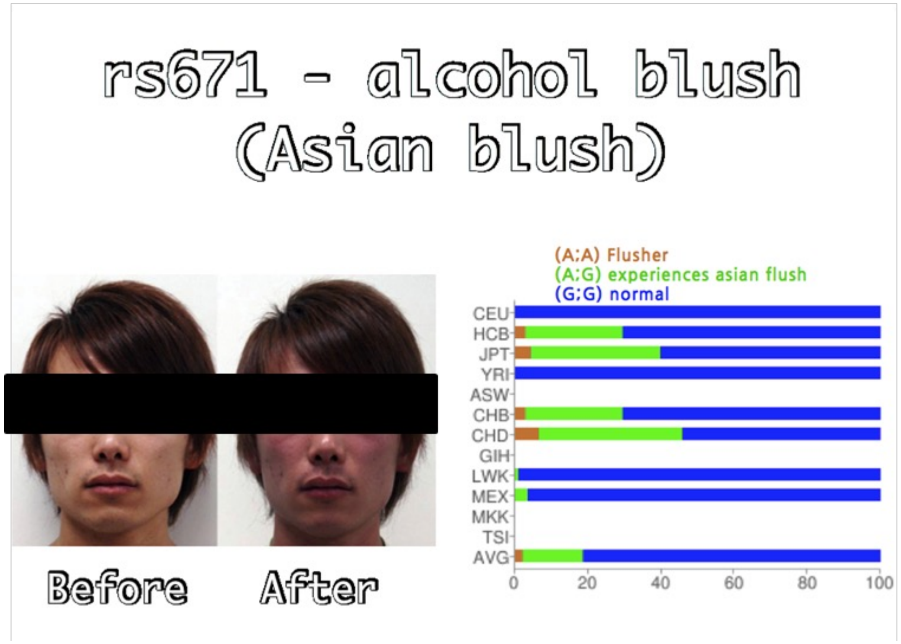
rs1234

HGP	ATCTAAGGCTATGAGGGCCCAATCTAGCTTTTATTTCCGAGAGA
토마스	ATCTAAGGCTATGAGGGCCCAATCTAGCTTTTATTTCCGACAGA
제임스	ATCTAAGGCTATGAGGGCCCAATCAAGCTTTTATTTCCGACAGA
제인	ATCTAGGGCTATGAGGGCCCAATCAAGCTTTTATTTCCGACAGA

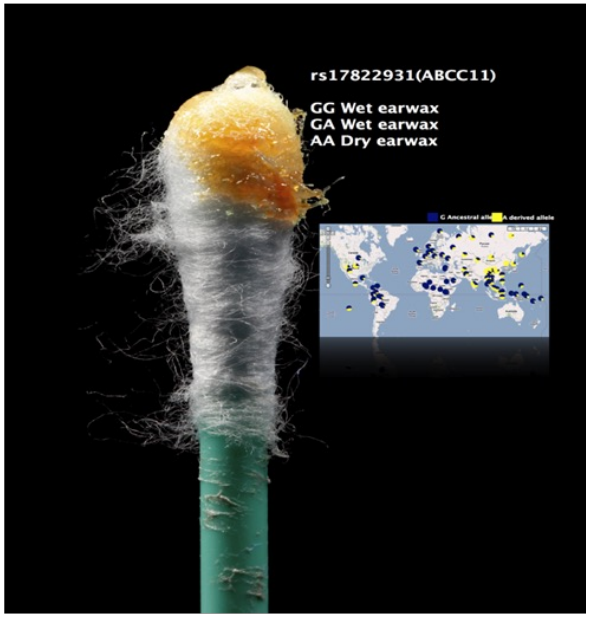
Single Nucleotide Polymorphism (SNP)

표현형의 차이




rs671 - alcohol blush (Asian blush)

[redacted]


rs17822931(ABCC11)

GG Wet earwax
GA Wet earwax
AA Dry earwax

Genetic diversity among populations

PCA analysis of East Asian descent

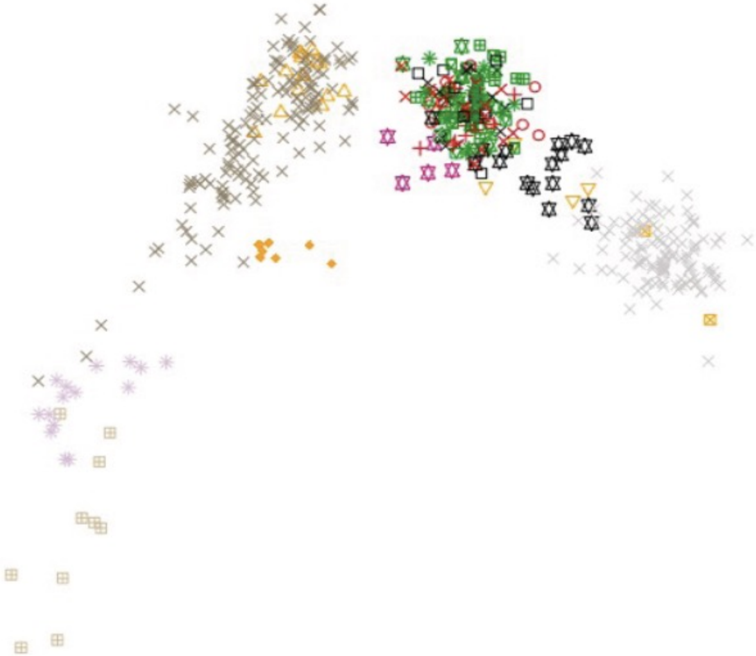
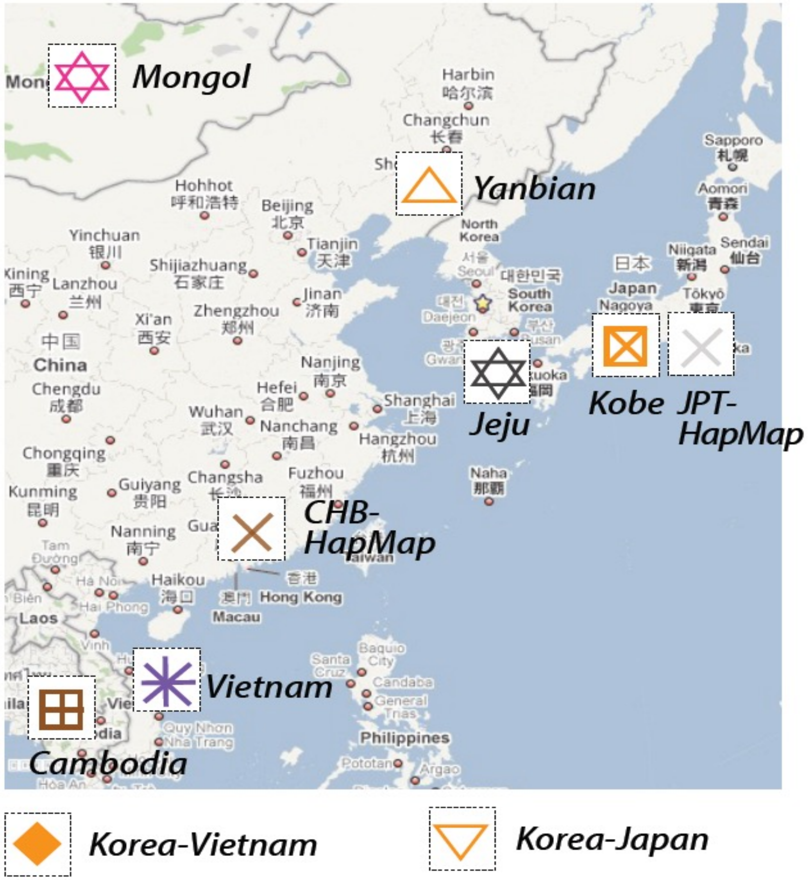


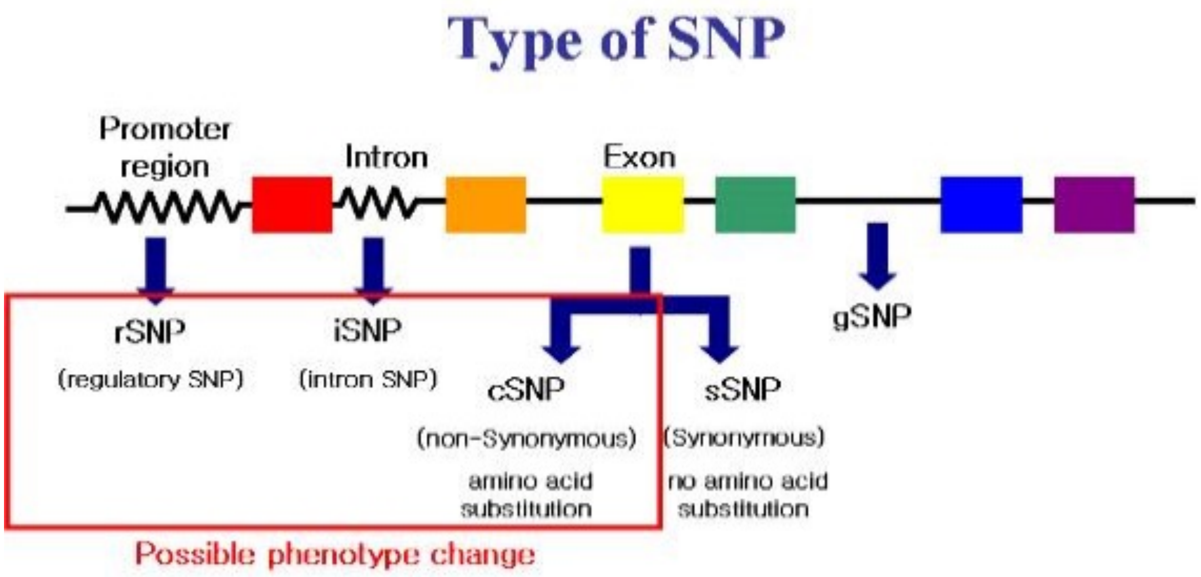
illustration of geographic correspondence of ethnic group locations



SNPs

- Synonymous: do not result in a change of amino acid in the protein, but still can affect its function in other ways

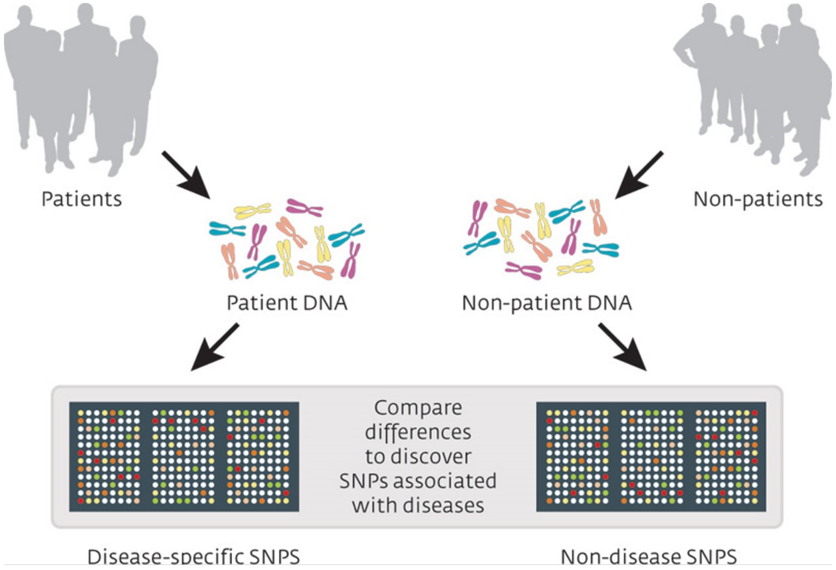
- Non-synonymous
  - Missense : amino acid changes
  - Nonsense : changes amino acid to stop codon



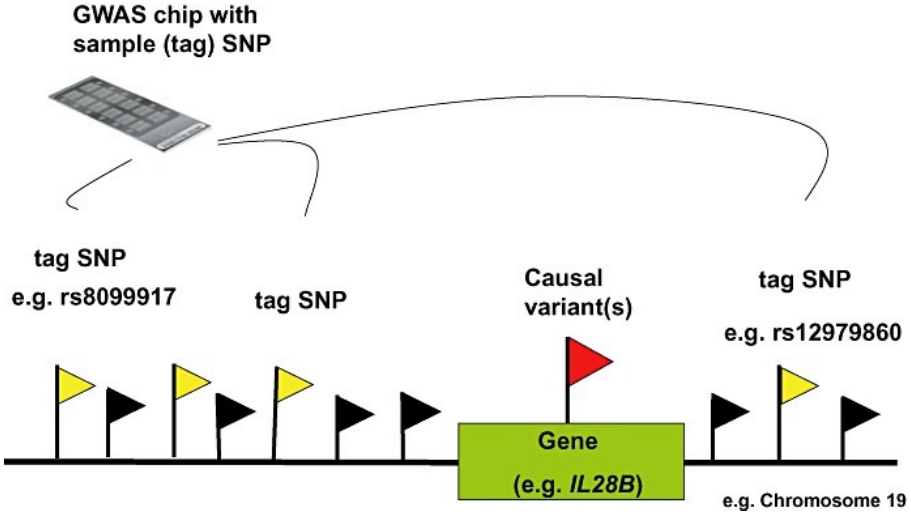
Genome-Wide Association Study (GWAS)

Genome-Wide Association Study (GWAS)

- Data driven, Hypothesis free
- Large # of variants across genome
- Exploratory

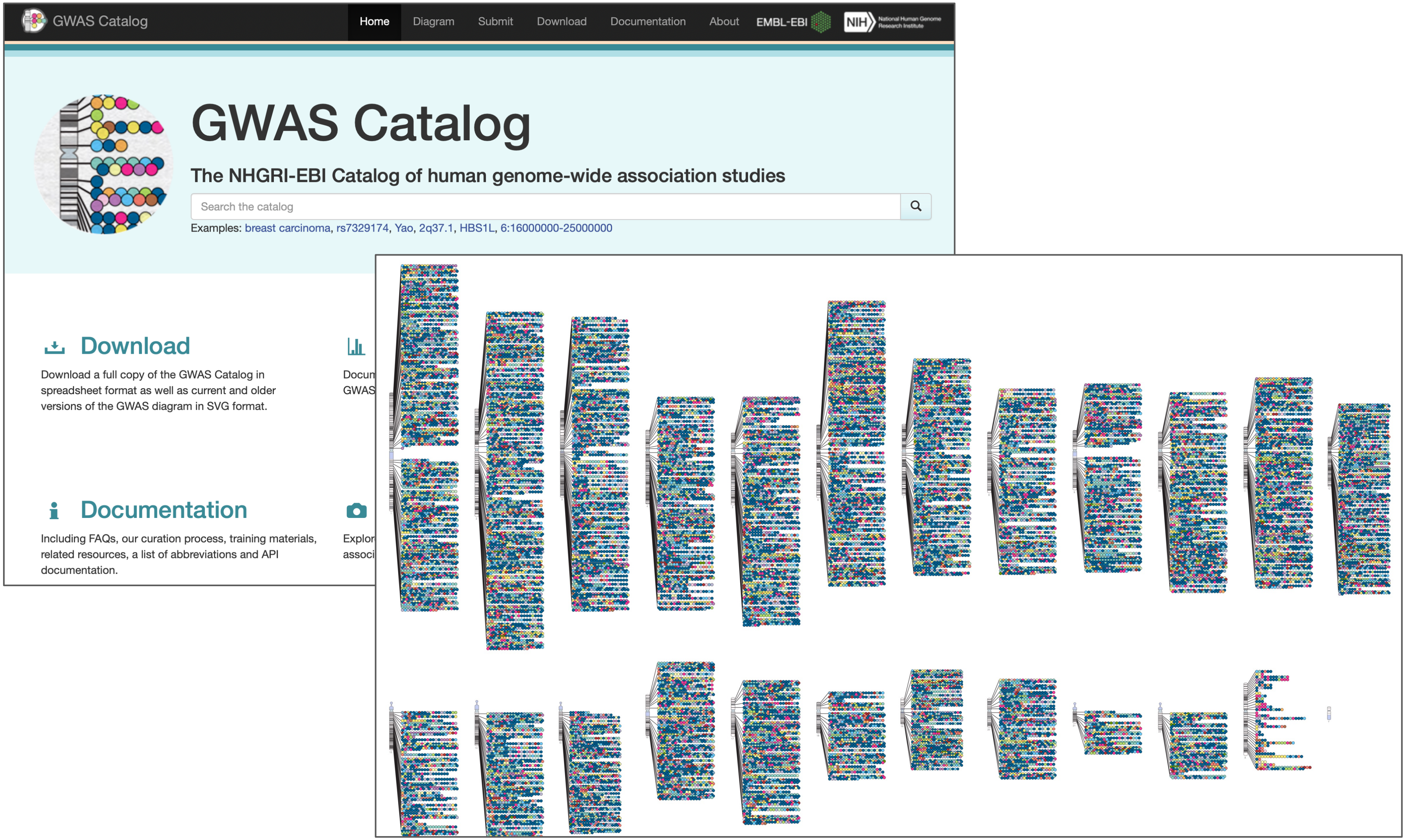


© Pasieka, Science Photo Library



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

Genetic variance VS. Disease risk



GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

Docun
GWAS

Documentation

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

Explor
associ

국가별 자국민 유전체칩 현황

국가	연도(시작)	샘플 수	사업명
미국	2011년	10만 명	UCSF-Kaiser RPGEH study
대만	2012년	10만 명	Taiwan Biobank Academia Sinica
영국	2013년	50만 명	UK Biobank
미국, 유럽 등	2013년	10만 명	iGeneTRAiN
미국	2013년	~100만 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

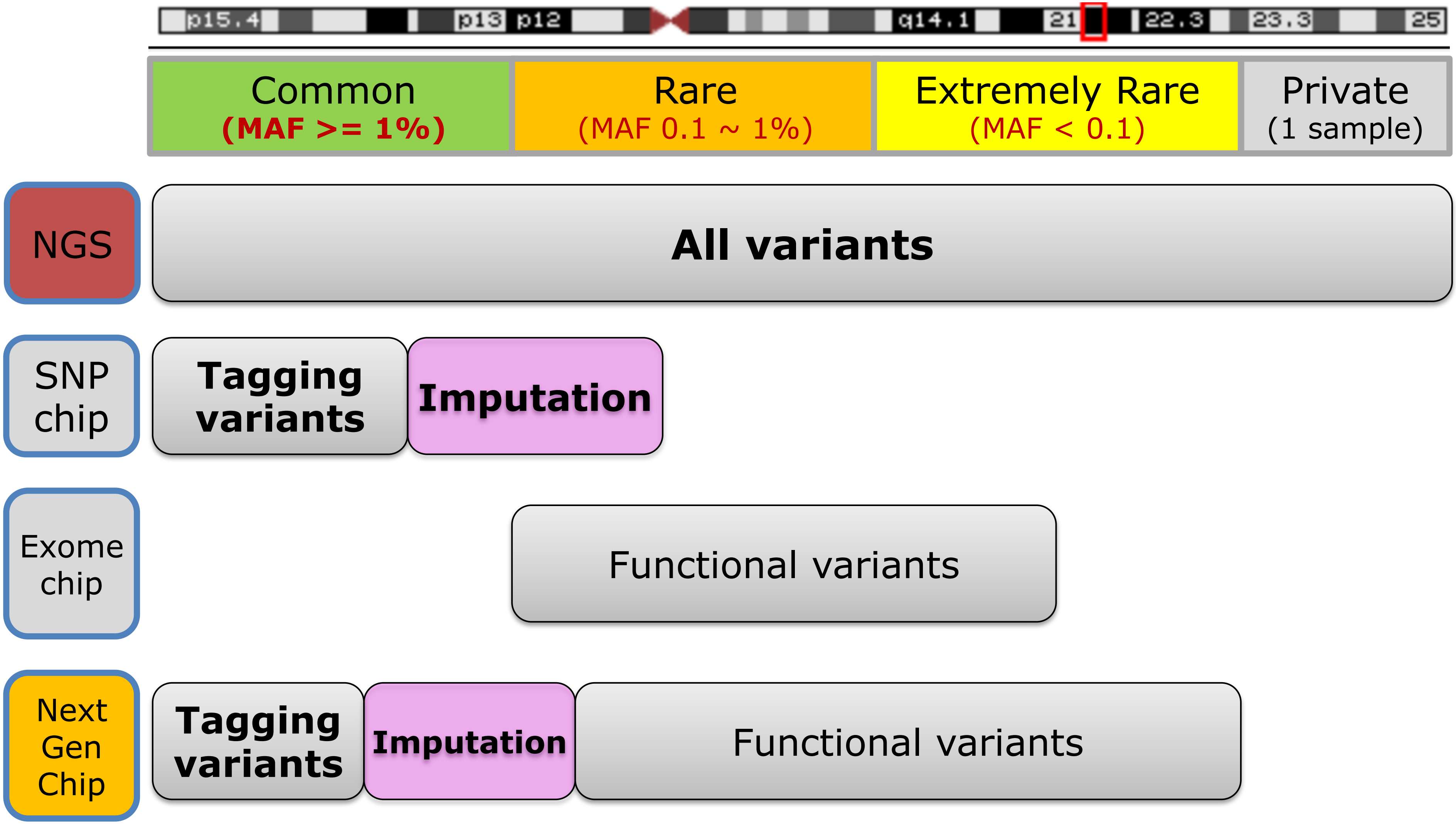
출처: Affymetrix

차세대 유전체칩

Arrays	Target Diseases	Main purpose	# of Contents	GW Tagging	Description
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
Global screening array	Complex Diseases	Discovery	700K	Yes	-
Global diversity array	Complex Diseases	Discovery	1.8M	Yes	All of Us project 300K

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

Strategies using Genome analysis technology



Common (MAF >= 1%)

Rare (MAF 0.1 ~ 1%)

Extremely Rare (MAF < 0.1)

Private (1 sample)

NGS: **All variants**

SNP chip: **Tagging variants**, **Imputation**

Exome chip: Functional variants

Next Gen Chip: **Tagging variants**, **Imputation**, Functional variants

한국인칩 제작

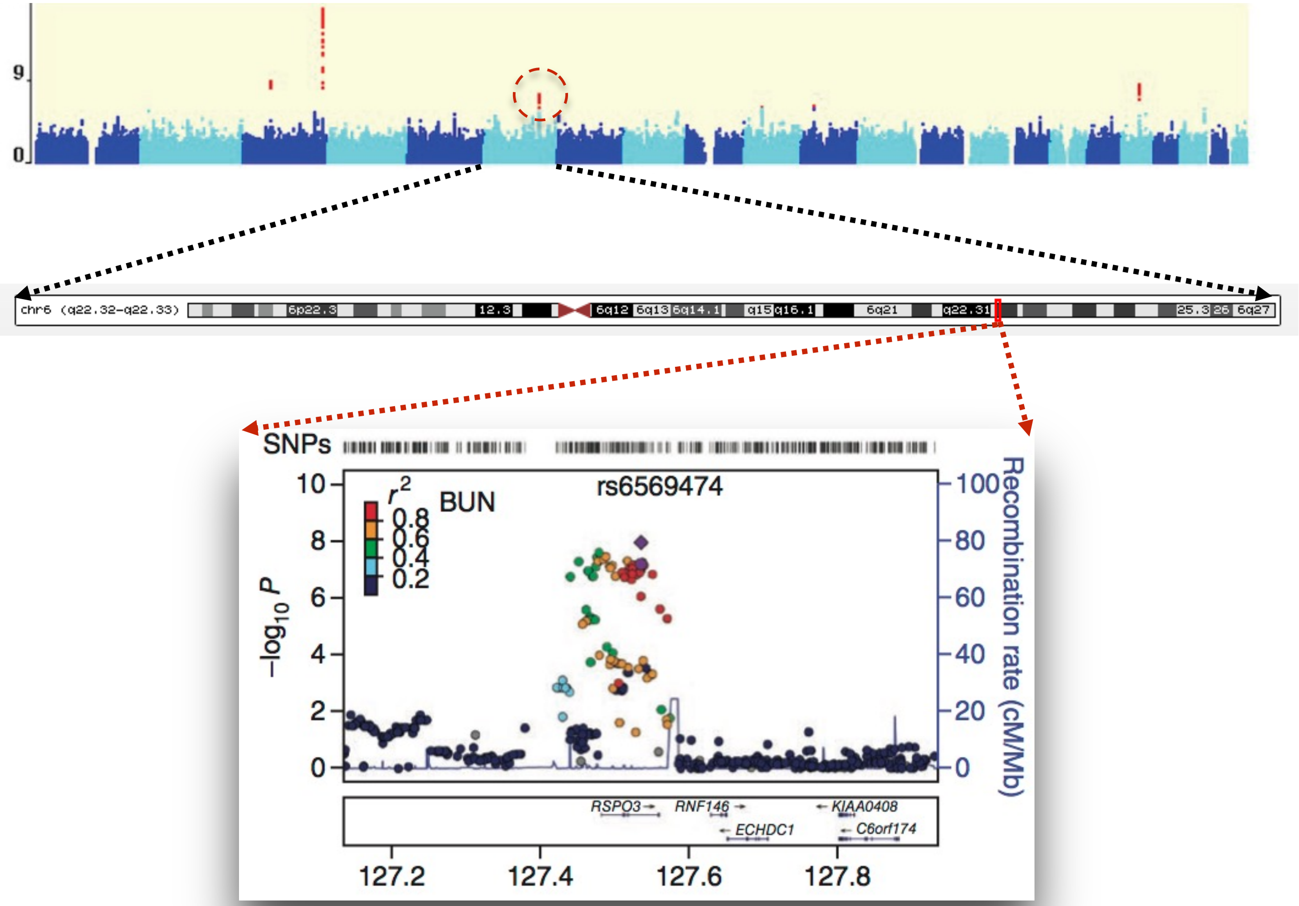
한국인칩 제작배경 : 기존 연구기법의 한계

- 한국인 만성질환 유전체 연구를 위한 대규모 인구집단 유전체 연구의 기존 연구기법의 문제점 대두
  - 유전체칩: 서양인 중심 설계, 한국인 염기서열정보 미반영
    - *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)

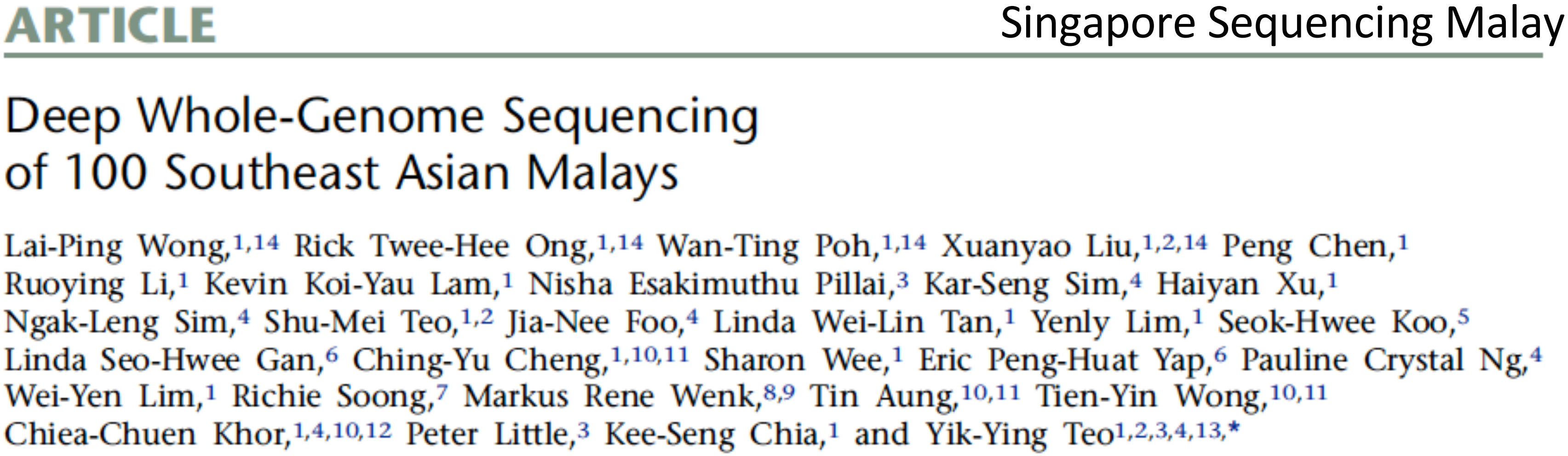
Example of Genome-wide scan

- High genomic coverage confers high association mapping power



Kim *et al.* Nature 2011

Singapore Sequencing Malay Project



ARTICLE

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

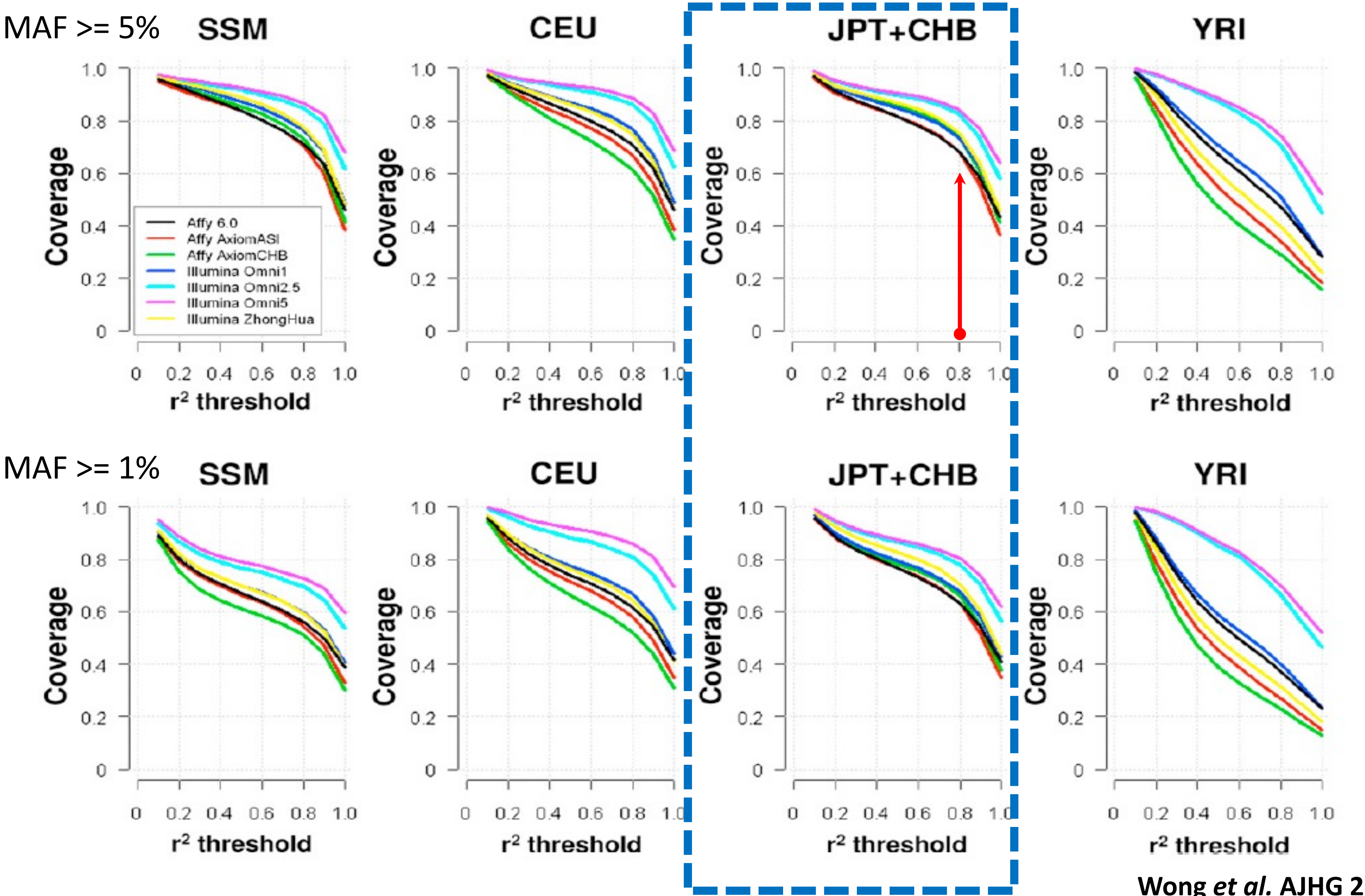
$$\text{Genomic Coverage} = \frac{\text{\# of Tagged markers}}{\text{Total \# of SNP}}$$

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



MAF >= 5%

MAF >= 1%

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

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)

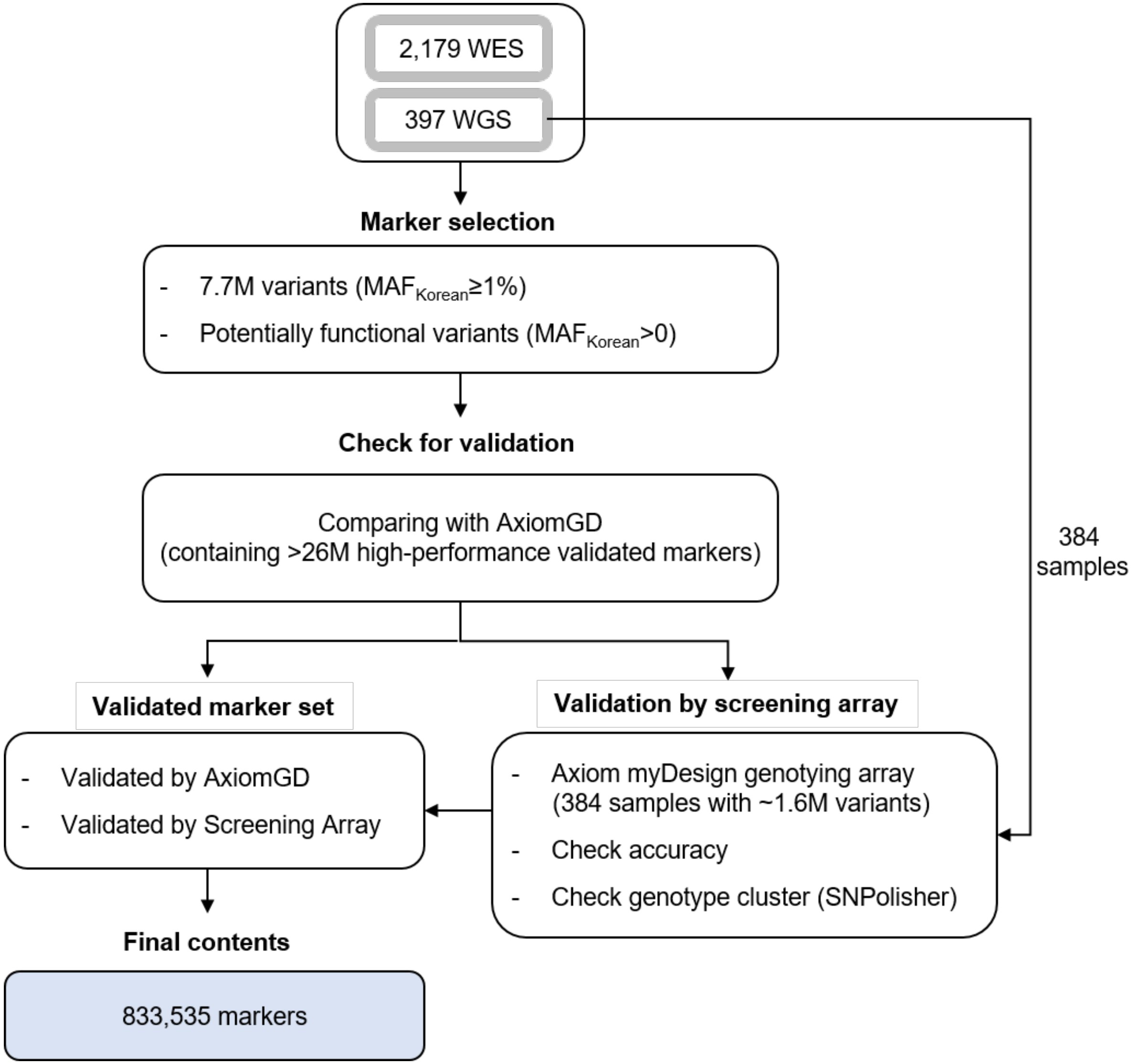
🟨 Axiom PMRA (96)

🟡 ILMN GSA (96)

CAVAS study

🟩 ILMN Omin 1 (3,666)

마커 선별 방법 (2014. 5 ~ 8)



2,179 WES

397 WGS

Marker selection

- 7.7M variants ($MAF_{Korean} \geq 1\%$)
- Potentially functional variants ($MAF_{Korean} > 0$)

Check for validation

Comparing with AxiomGD
(containing >26M high-performance validated markers)

384 samples

Validated marker set

- Validated by AxiomGD
- Validated by Screening Array

Validation by screening array

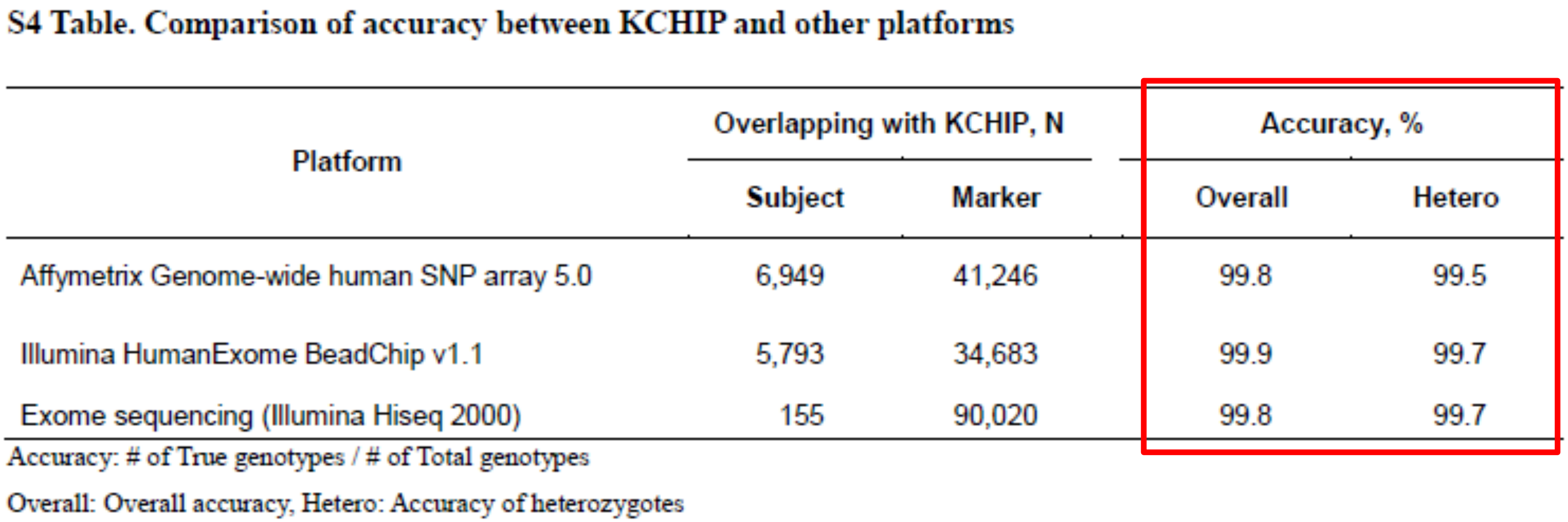
- Axiom myDesign genotyping array (384 samples with ~1.6M variants)
- Check accuracy
- Check genotype cluster (SNPolisher)

Final contents

833,535 markers

정확도

Methods-Accuracy and reproducibility



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

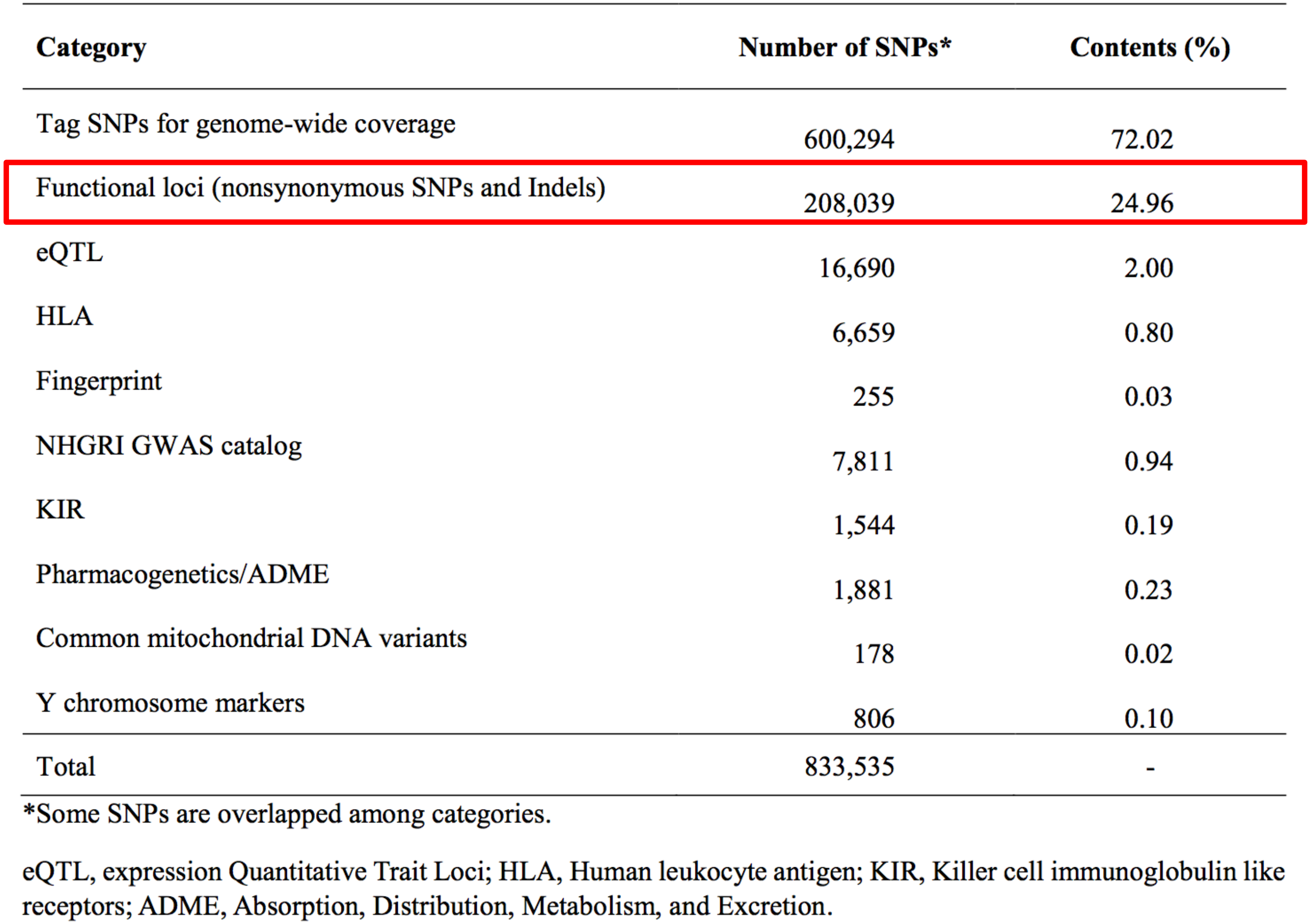
Overall: Overall accuracy, Hetero: Accuracy of heterozygotes

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

콘텐츠

한국인칩 콘텐츠

Table 1. Contents summary of KoreanChip

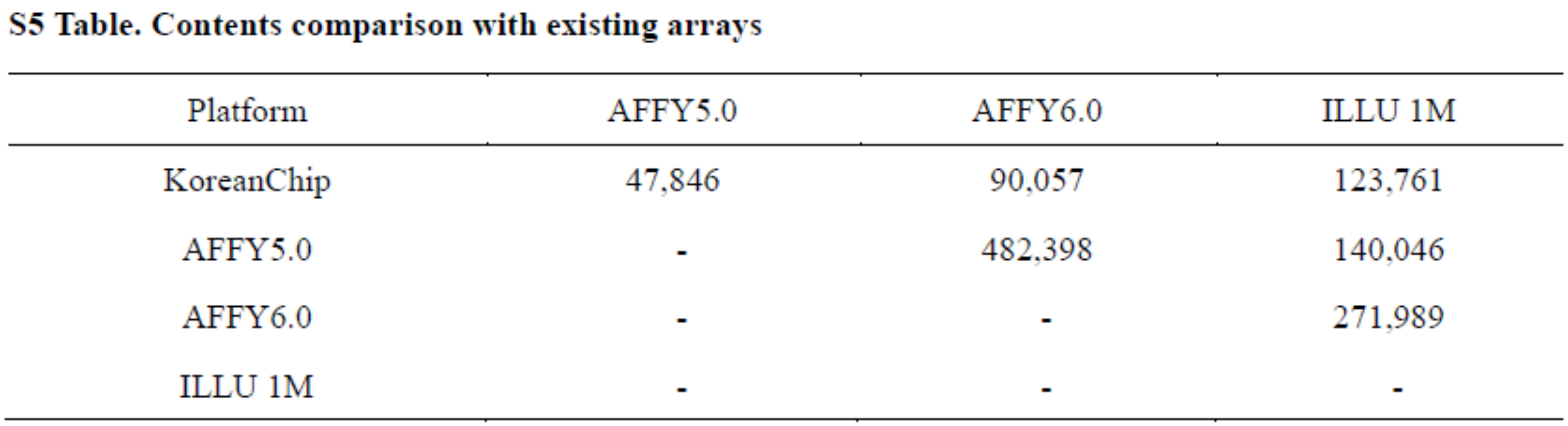


Category	Number of SNPs*	Contents (%)
Tag SNPs for genome-wide coverage	600,294	72.02
Functional loci (nonsynonymous SNPs and Indels)	208,039	24.96
eQTL	16,690	2.00
HLA	6,659	0.80
Fingerprint	255	0.03
NHGRI GWAS catalog	7,811	0.94
KIR	1,544	0.19
Pharmacogenetics/ADME	1,881	0.23
Common mitochondrial DNA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

*Some SNPs are overlapped among categories.

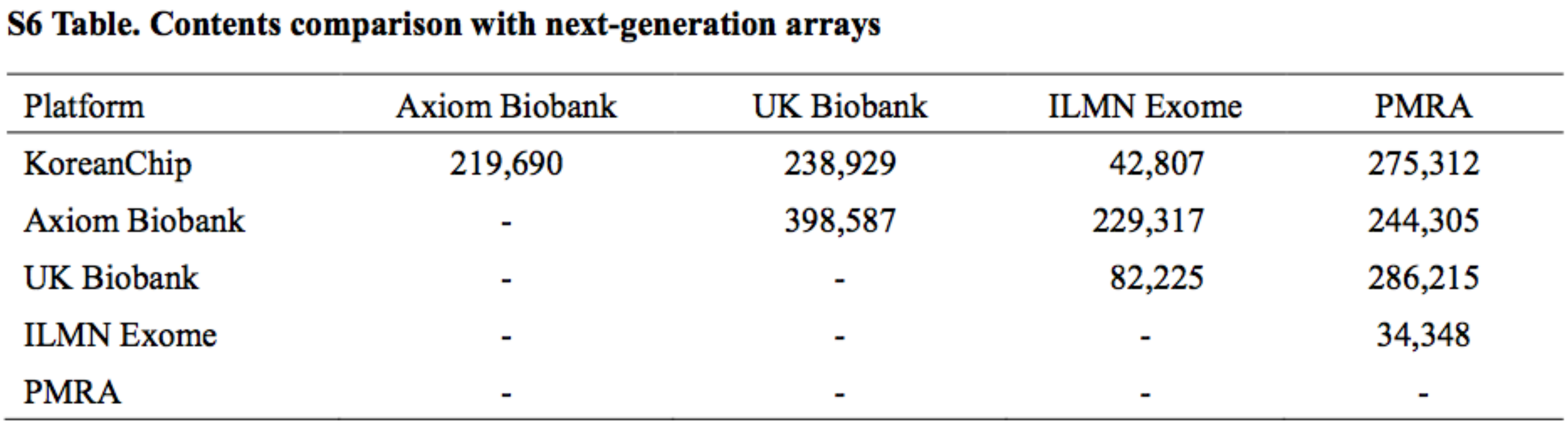
eQTL, expression Quantitative Trait Loci; HLA, Human leukocyte antigen; KIR, Killer cell immunoglobulin like receptors; ADME, Absorption, Distribution, Metabolism, and Excretion.

상용칩과 콘텐츠 비교



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

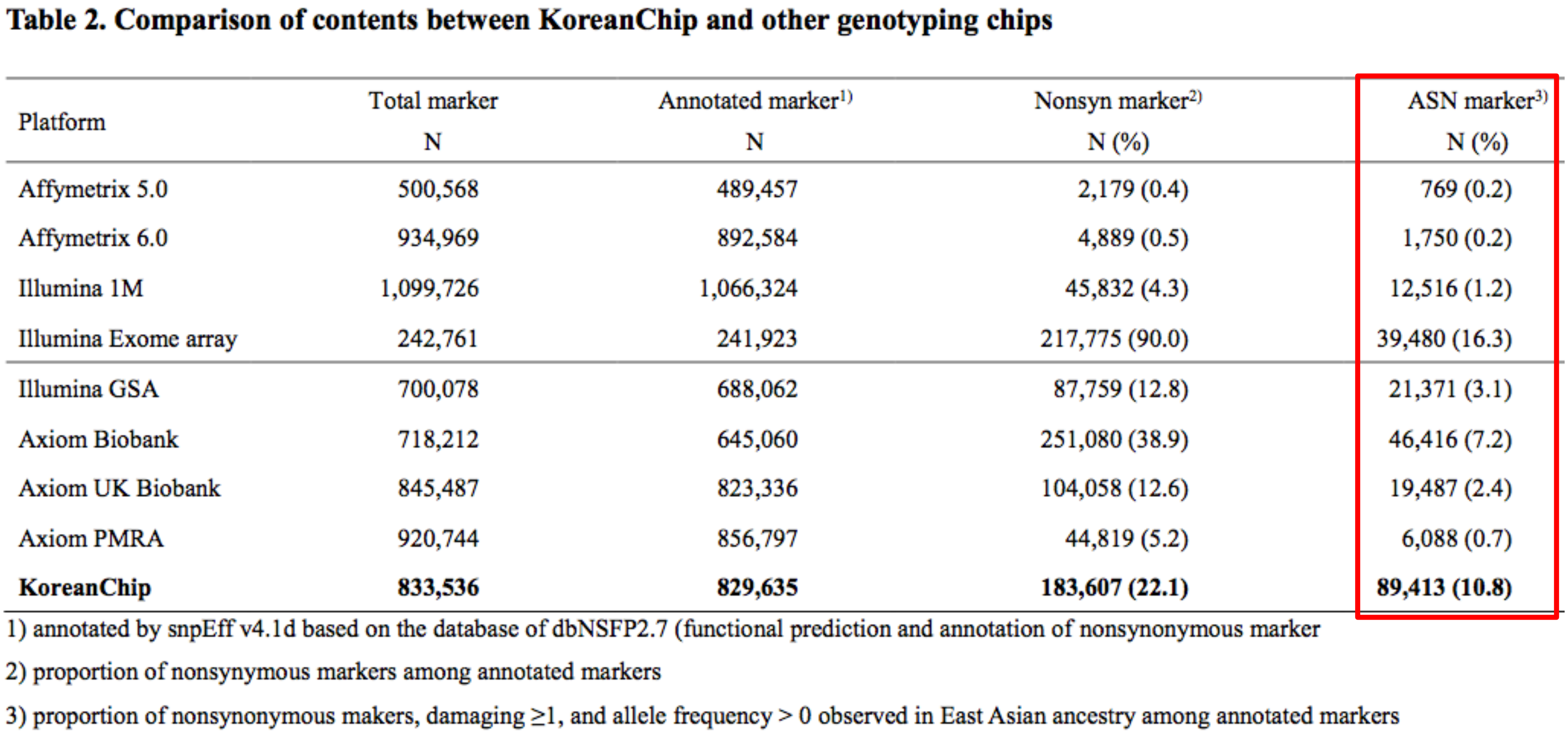


Table 2. Comparison of contents between KoreanChip and other genotyping chips

Platform	Total marker	Annotated marker[1]	Nonsyn marker[2]	ASN marker[3]
	N	N	N (%)	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

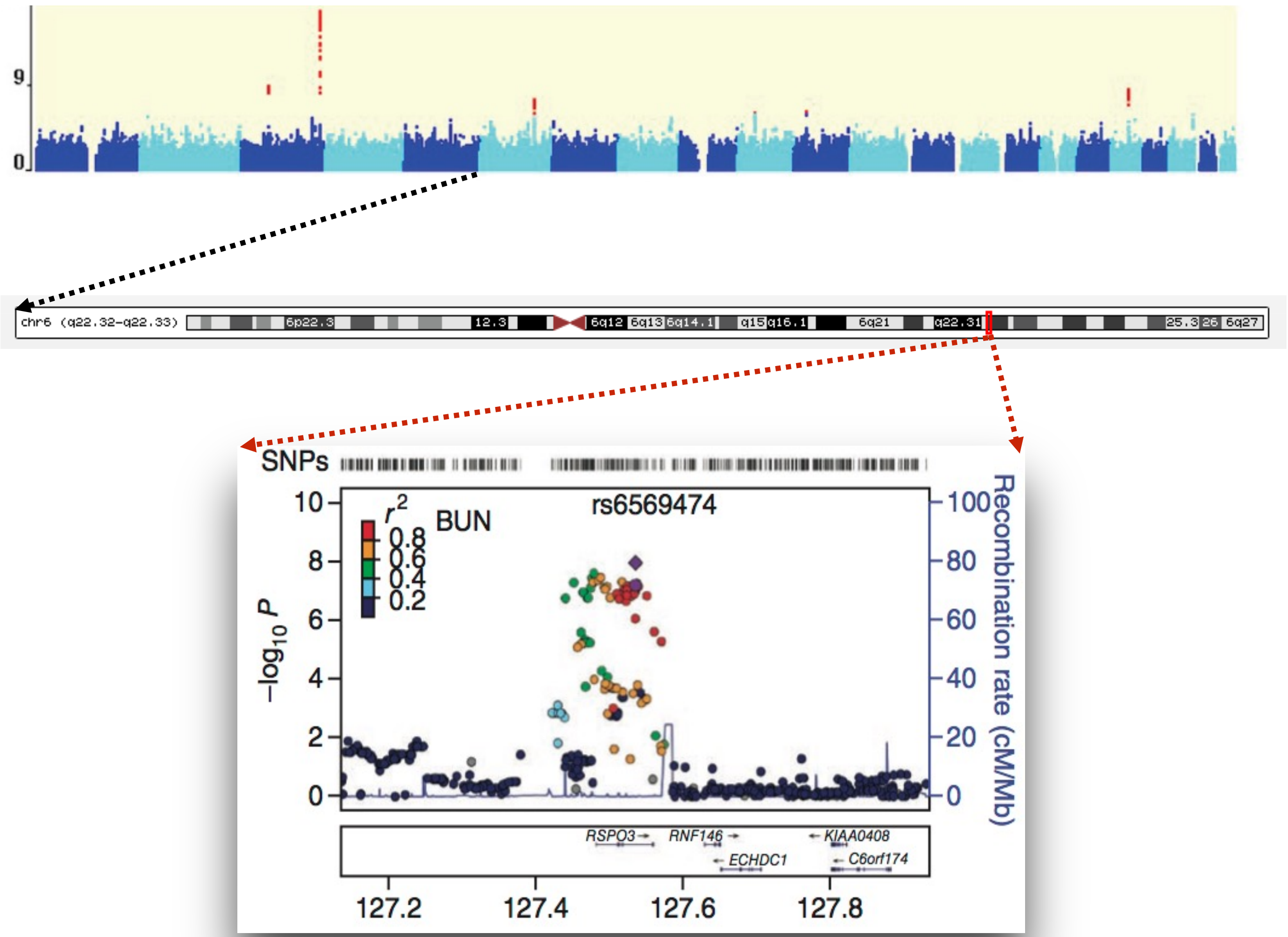
2) proportion of nonsynymous markers among annotated markers

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

Genomic coverage

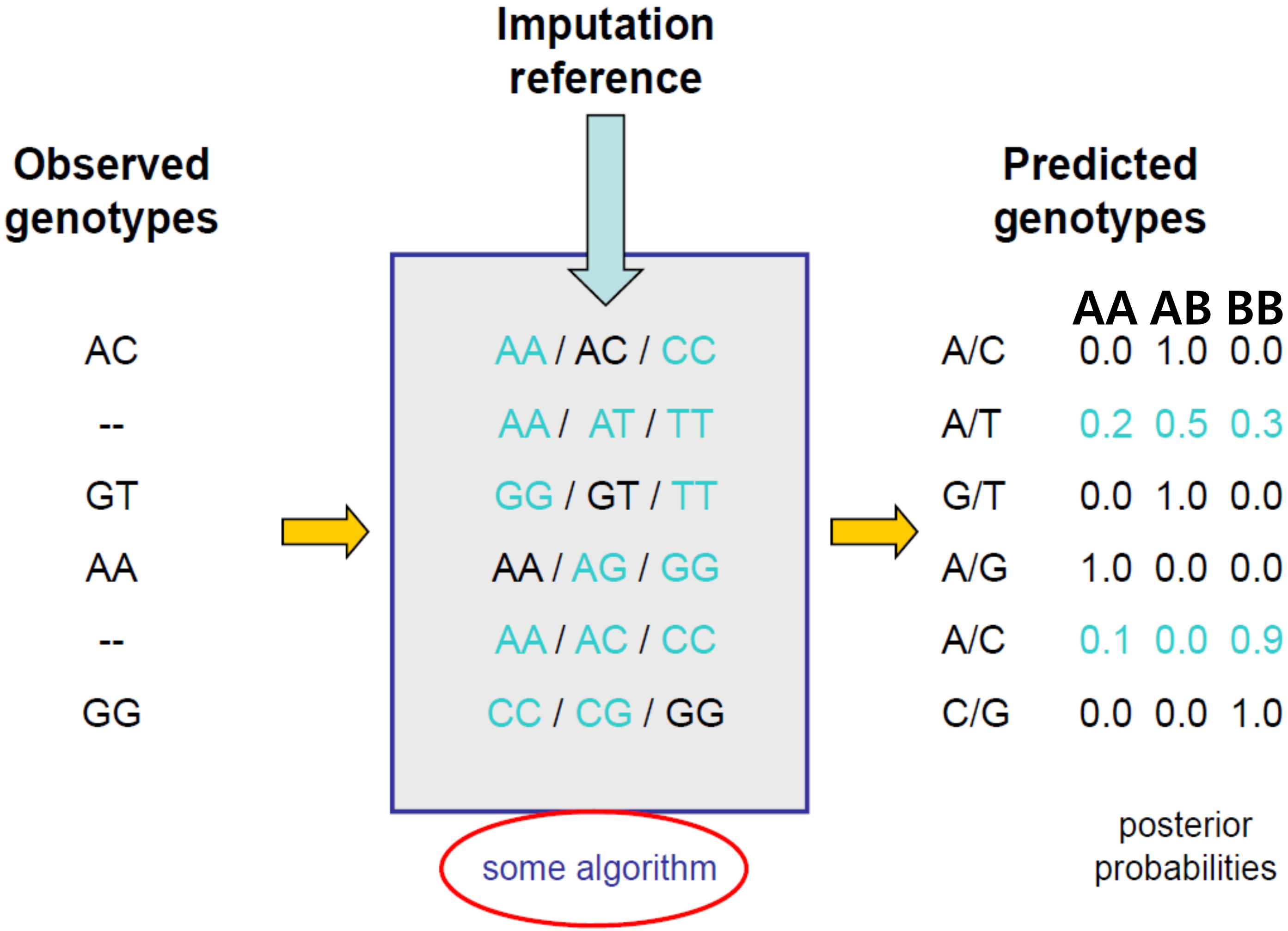
Example of Genome-wide scan

- High genomic coverage confers high association mapping power



Kim *et al.* Nature 2011

Introduction – Imputation



Imputation reference

Observed genotypes

AC
--
GT
AA
--
GG

AA / AC / CC
AA / AT / TT
GG / GT / TT
AA / AG / GG
AA / AC / CC
CC / CG / GG

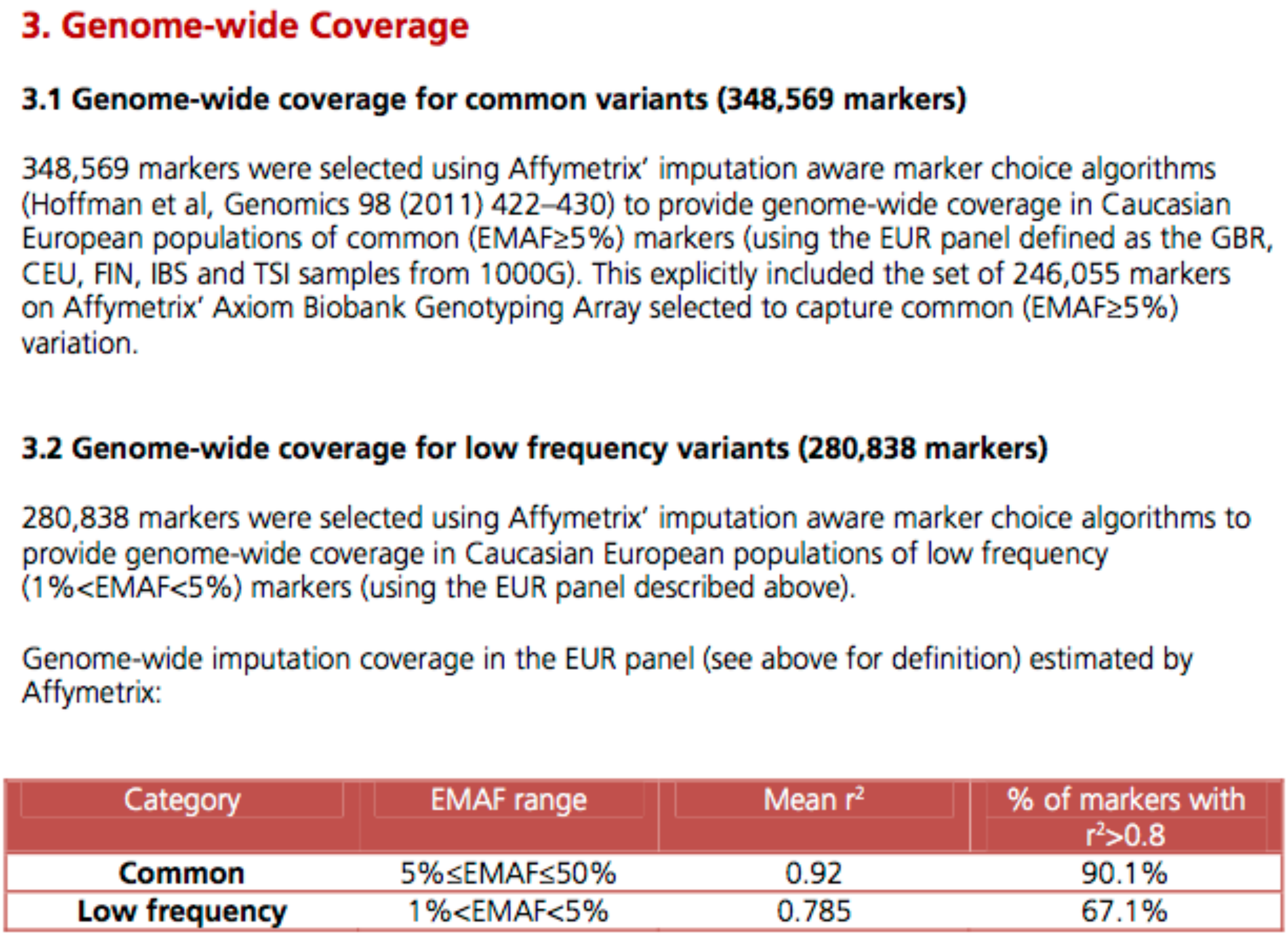
some algorithm

Predicted genotypes

	AA	AB	BB
A/C	0.0	1.0	0.0
A/T	0.2	0.5	0.3
G/T	0.0	1.0	0.0
A/G	1.0	0.0	0.0
A/C	0.1	0.0	0.9
C/G	0.0	0.0	1.0

posterior
probabilities

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

Estimated genomic coverage

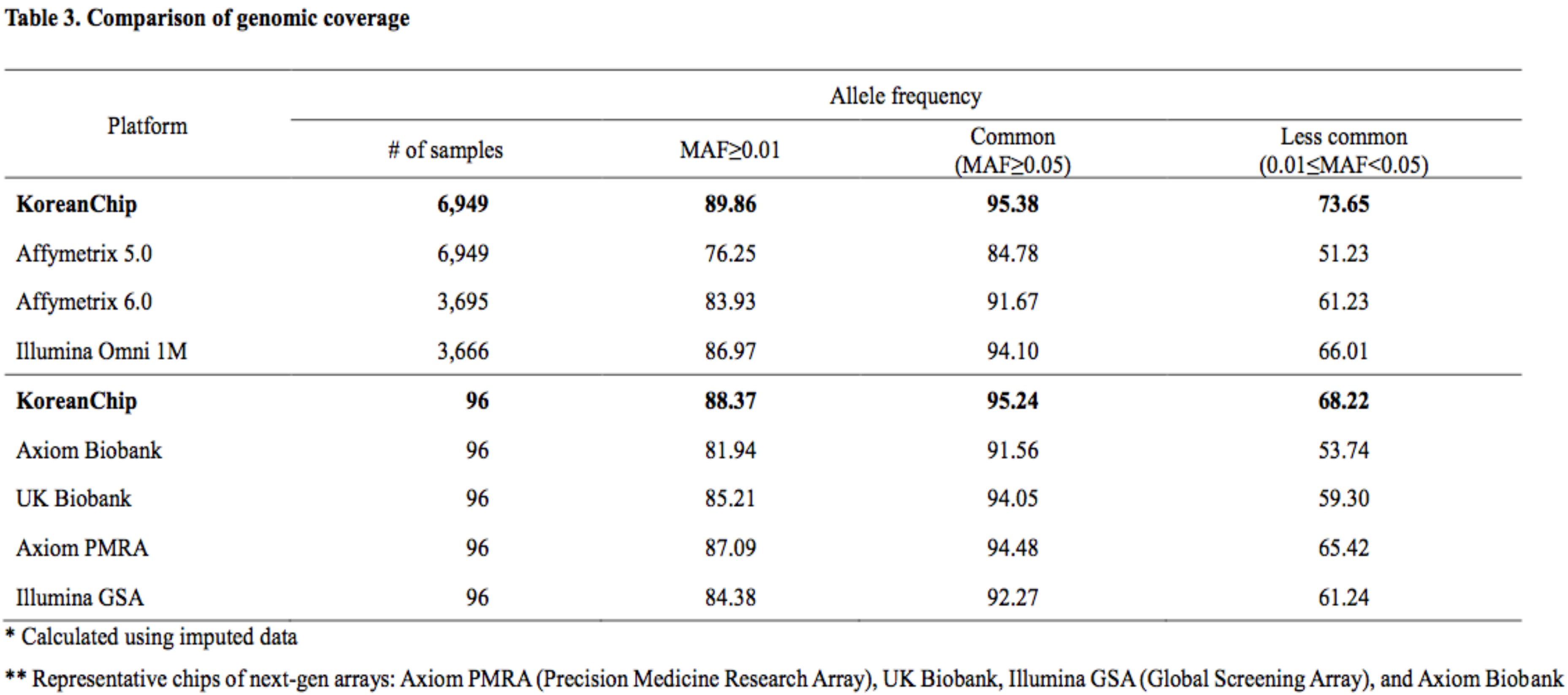


Table 3. Comparison of genomic coverage

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

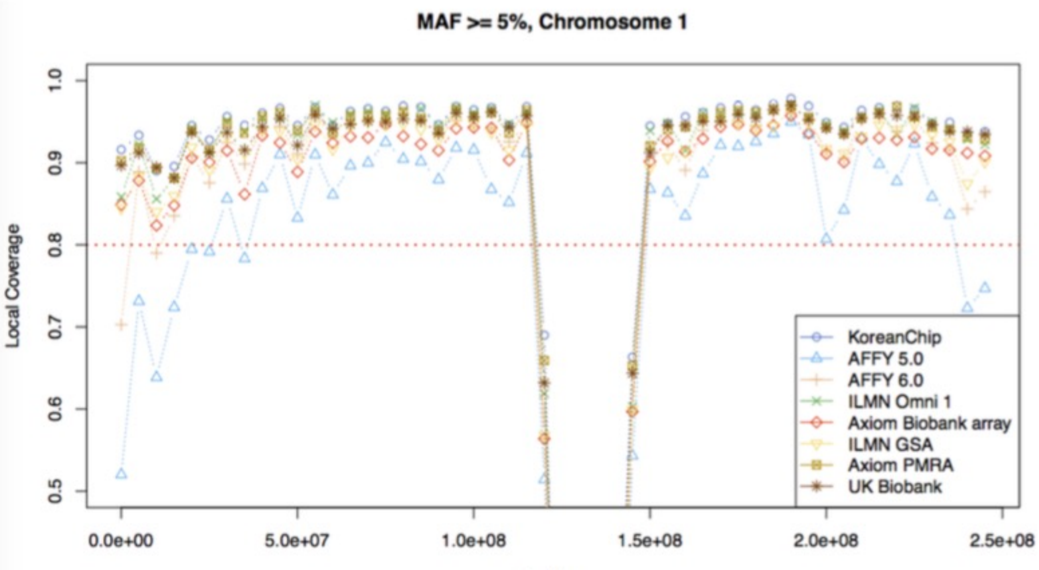
* Calculated using imputed data

** 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

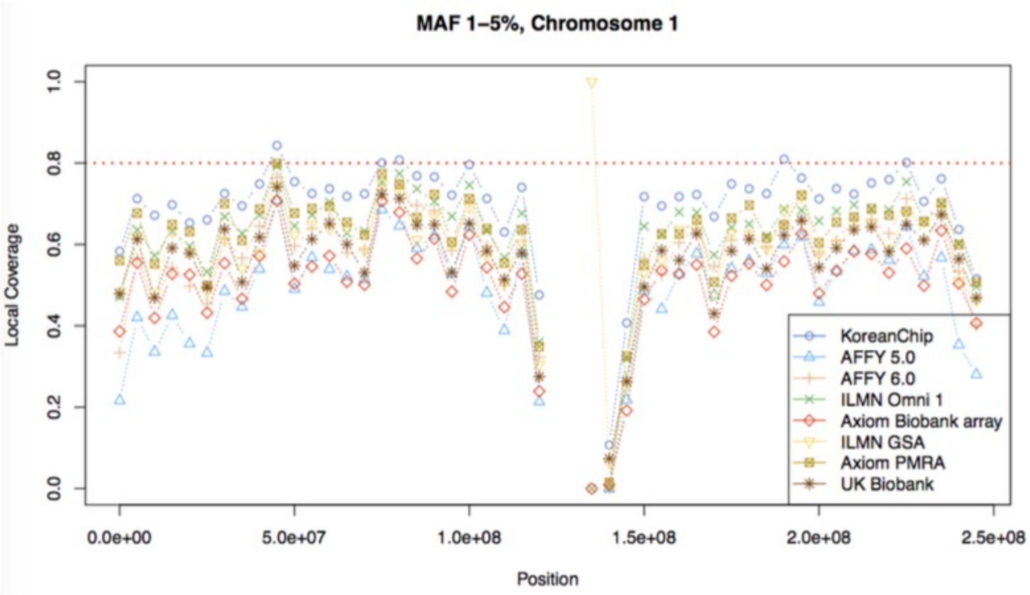
CHR 1

(MAF ≥ 5%)

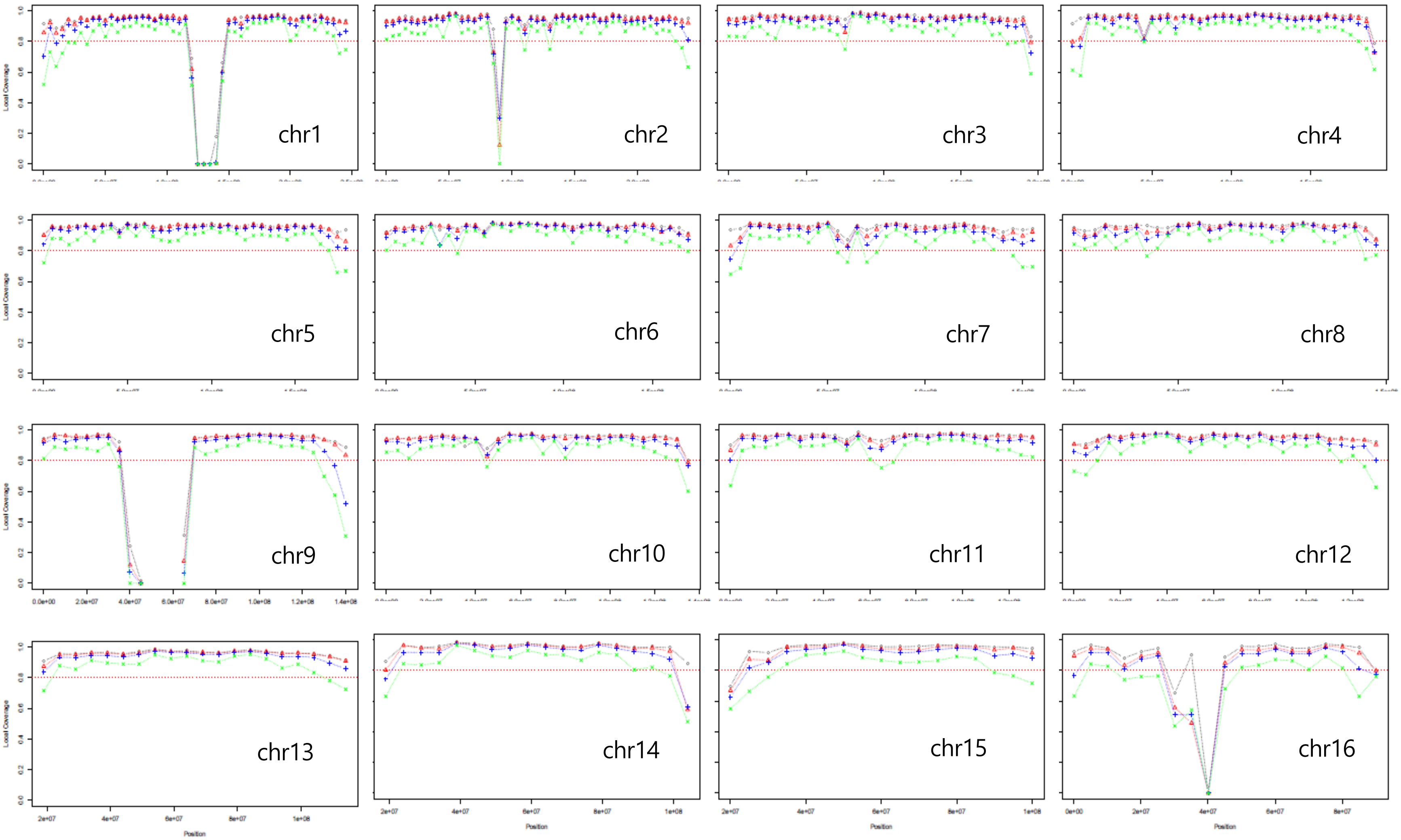


CHR 1

(MAF 1-5%)

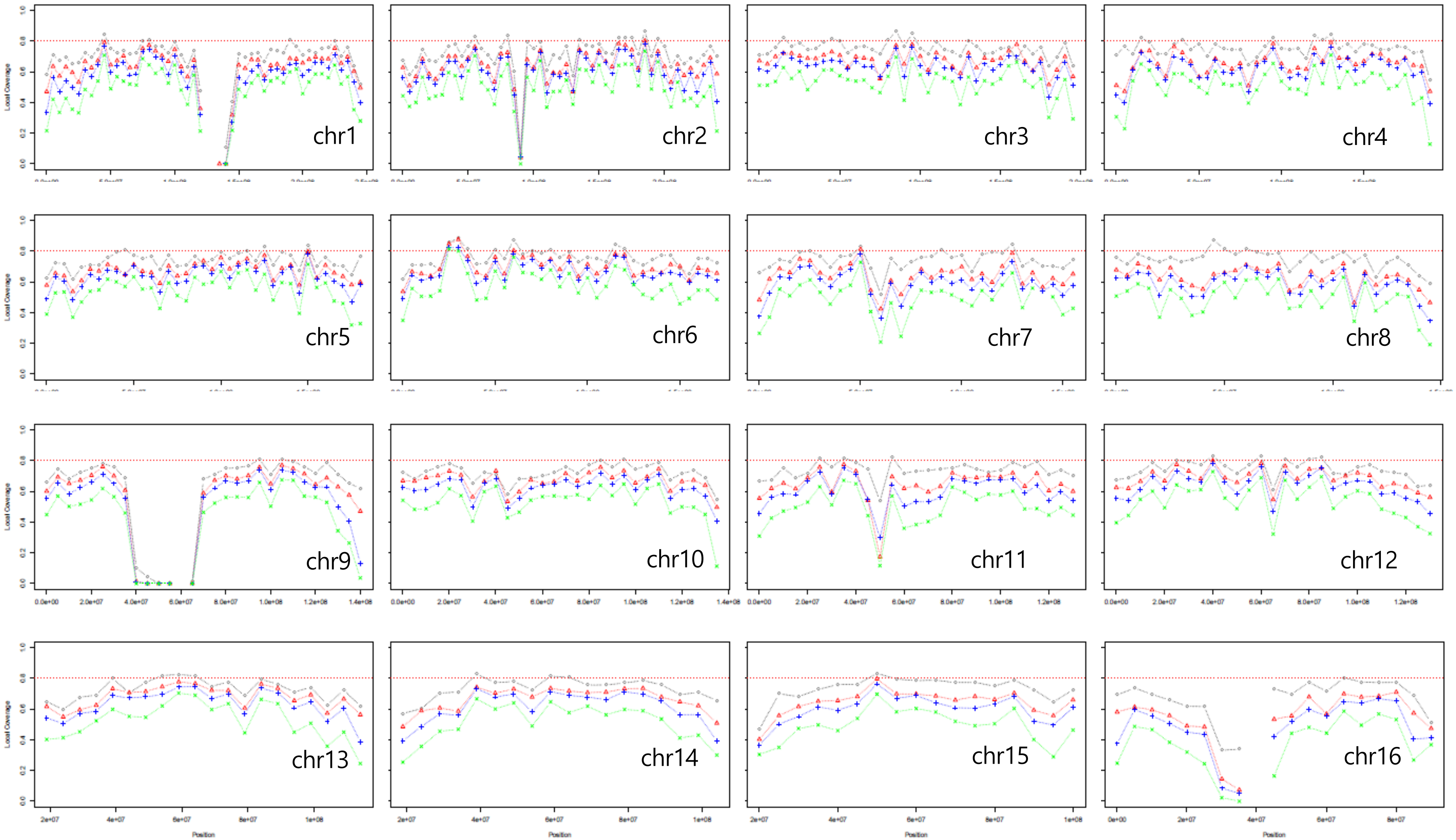


Estimated genomic coverage(MAF ≥ 5%)



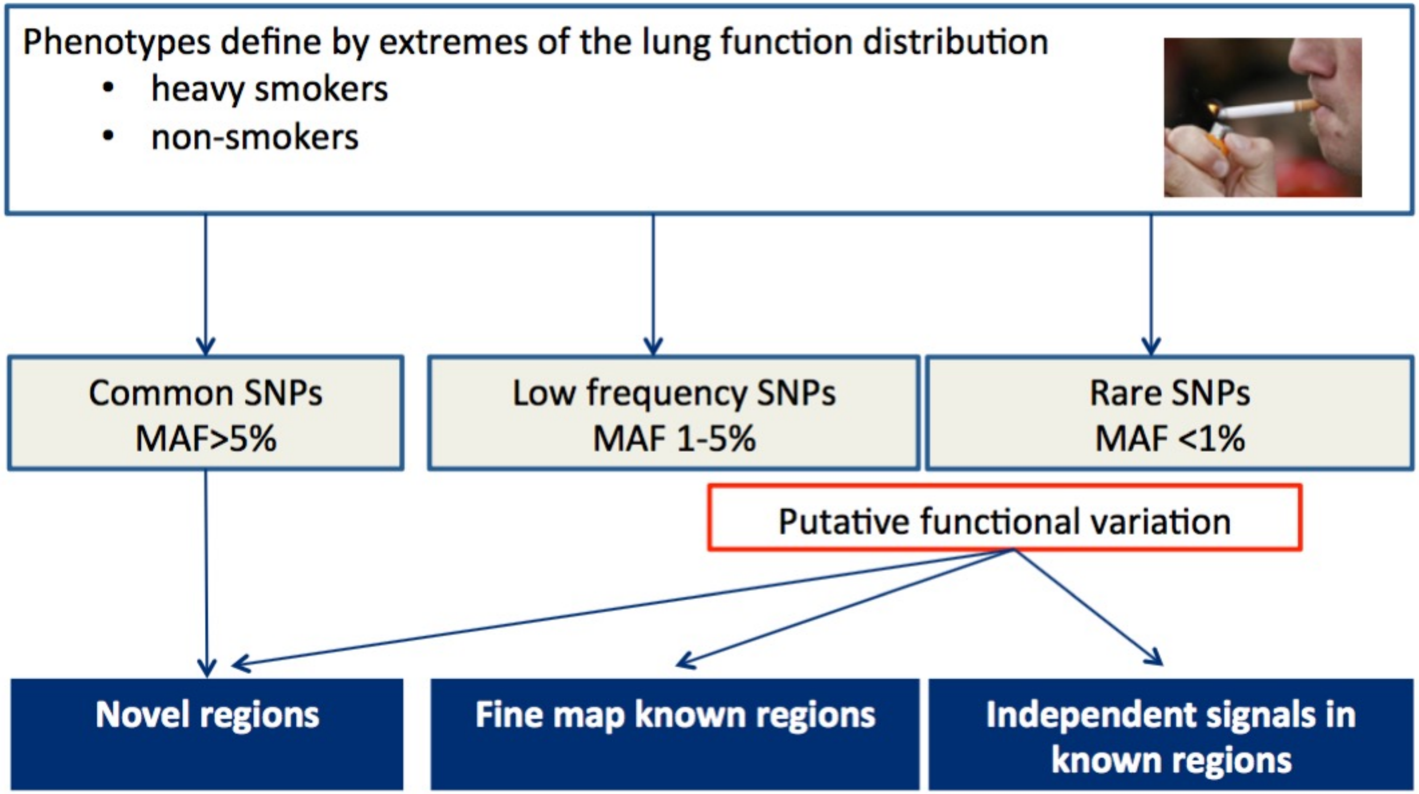
Estimated genomic coverage(MAF 1-5%)

국립보건연구원
National Institute of Health



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%

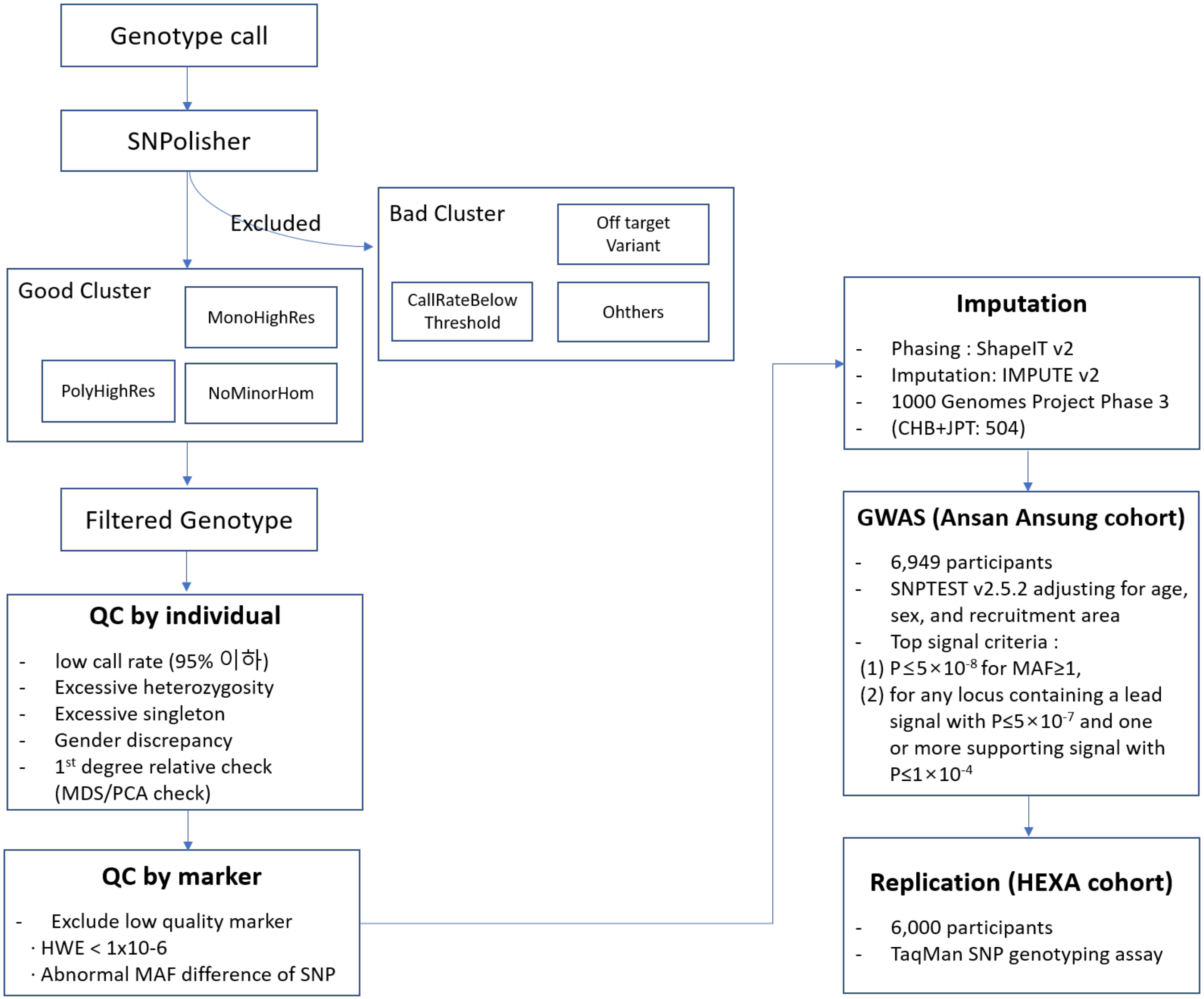
Putative functional variation

Novel regions

Fine map known regions

Independent signals in known regions

Overall scheme of GWAS



Genotype call

SNPolisher

Excluded

Bad Cluster

Off target Variant

CallRateBelow Threshold

Ohthers

Good Cluster

MonoHighRes

PolyHighRes

NoMinorHom

Filtered Genotype

QC by individual

- low call rate (95% 이하)
- Excessive heterozygosity
- Excessive singleton
- Gender discrepancy
- 1st degree relative check (MDS/PCA check)

QC by marker

- Exclude low quality marker
  · HWE < 1x10-6
  · Abnormal MAF difference of SNP

Imputation

- Phasing : ShapeIT v2
- Imputation: IMPUTE v2
- 1000 Genomes Project Phase 3
- (CHB+JPT: 504)

GWAS (Ansan Ansung cohort)

- 6,949 participants
- SNPTEST v2.5.2 adjusting for age, sex, and recruitment area
- Top signal criteria :

(1) $P \le 5 \times 10^{-8}$ for MAF≥1,

(2) for any locus containing a lead signal with $P \le 5 \times 10^{-7}$ and one or more supporting signal with $P \le 1 \times 10^{-4}$

Replication (HEXA cohort)

- 6,000 participants
- TaqMan SNP genotyping assay

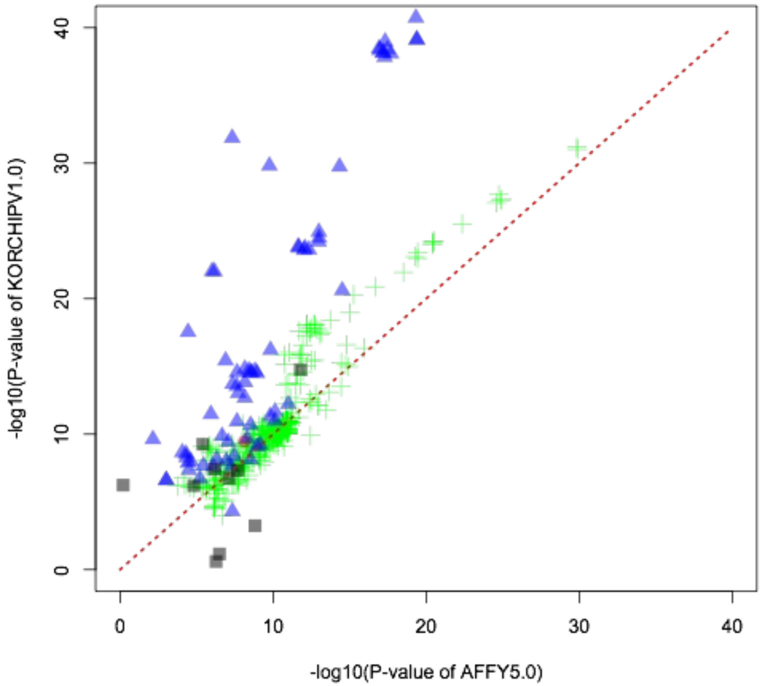
Comparison results of Association signals

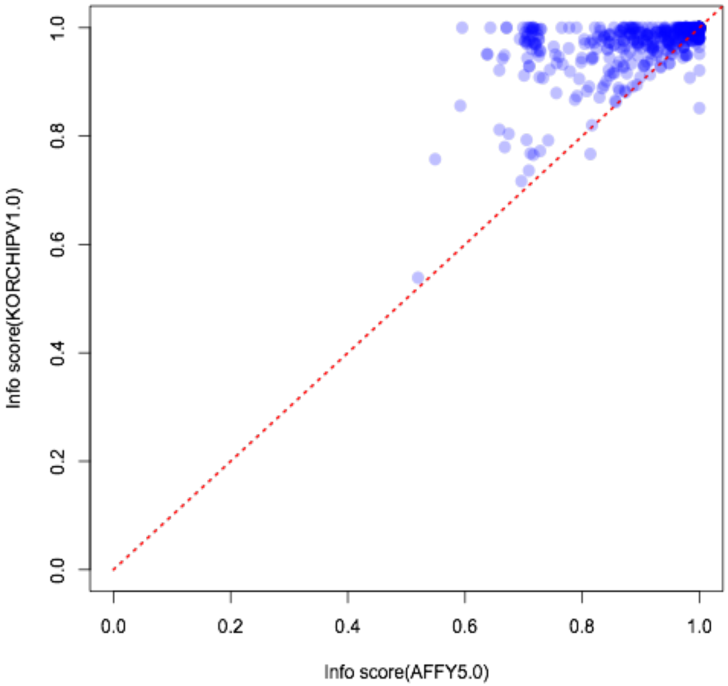
- Comparison analysis
  - Data: AFFY5.0, KORV1.0 identical 7,000 samples (Imputed using 1KG phase3, 8,700,150 variants)
  - Phenotype: Lipids (HDL, LDL, TG), Liver enzyme (AST, ALT, GGT), T2D
  - Association test: SNPTEST v2.5
  - Covariates: age, gender, recruitment area
  - Top signal selection
  -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

Comparison results of Association signals

국립보건연구원
National Institute of Health

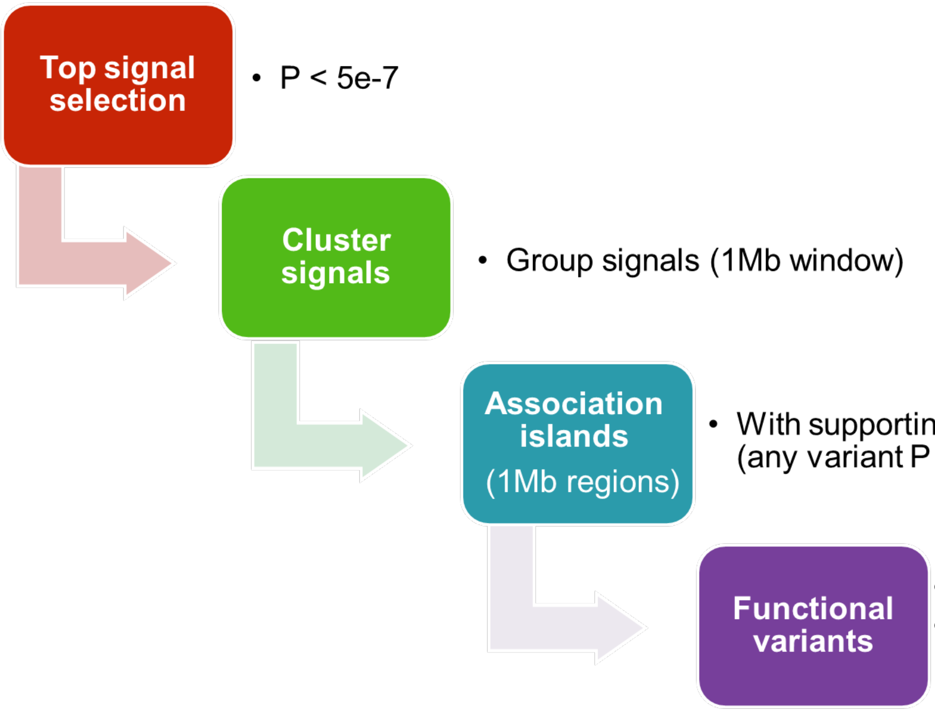
- Association results (HDL)
  - High quality (info score > 0.8): similar association results
  - In overall, K-CHIP showed higher imputation quality and stronger statistical significance





Preliminary association analysis

- 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)



Top signal selection

- P < 5e-7

Cluster signals

- Group signals (1Mb window)

Association islands (1Mb regions)

- With supporting evidence (any variant P < 5e-4)

Functional variants

- Genotyped
- Predicted to be damaging(dbNSFP)

31 variants remained

Application to GWAS (known or novel variants)

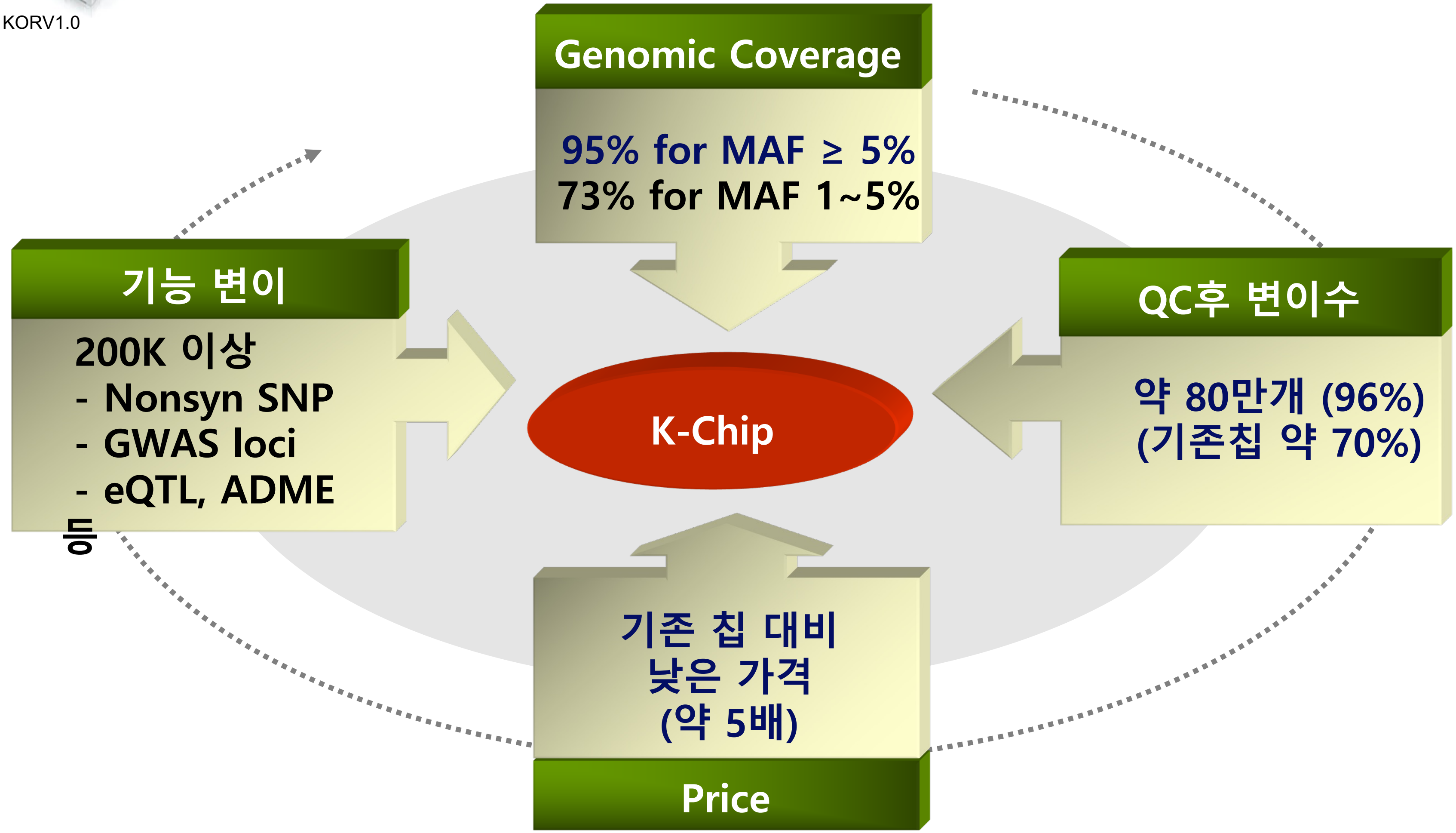
Gene	Trait(s)	EAF(%)				Discovery (~6,949 samples)		Replication (~6,000 samples)	
		KOR	gnomAD						
			EAS	EUR	AFR	Beta(SE)	P-value	Beta(SE)	P-value
5 variants at known loci									
-	TG	33.31	37.00	20.86	21.53	-0.0415(0.0089)	3.27E-06	-0.0483(0.0105)	4.26E-06
C2orf16	TG	52.87	47.81	27.16	6.61	0.0379(0.0084)	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
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

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

한국인칩 장점

국립보건연구원
National Institute of Health

KORV1.0



Genomic Coverage

95% for MAF ≥ 5%
73% for MAF 1~5%

기능 변이

200K 이상
- Nonsyn SNP
- GWAS loci
- eQTL, ADME
등

K-Chip

QC후 변이수

약 80만개 (96%)
(기존칩 약 70%)

기존 칩 대비
낮은 가격
(약 5배)

Price

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

감사합니다