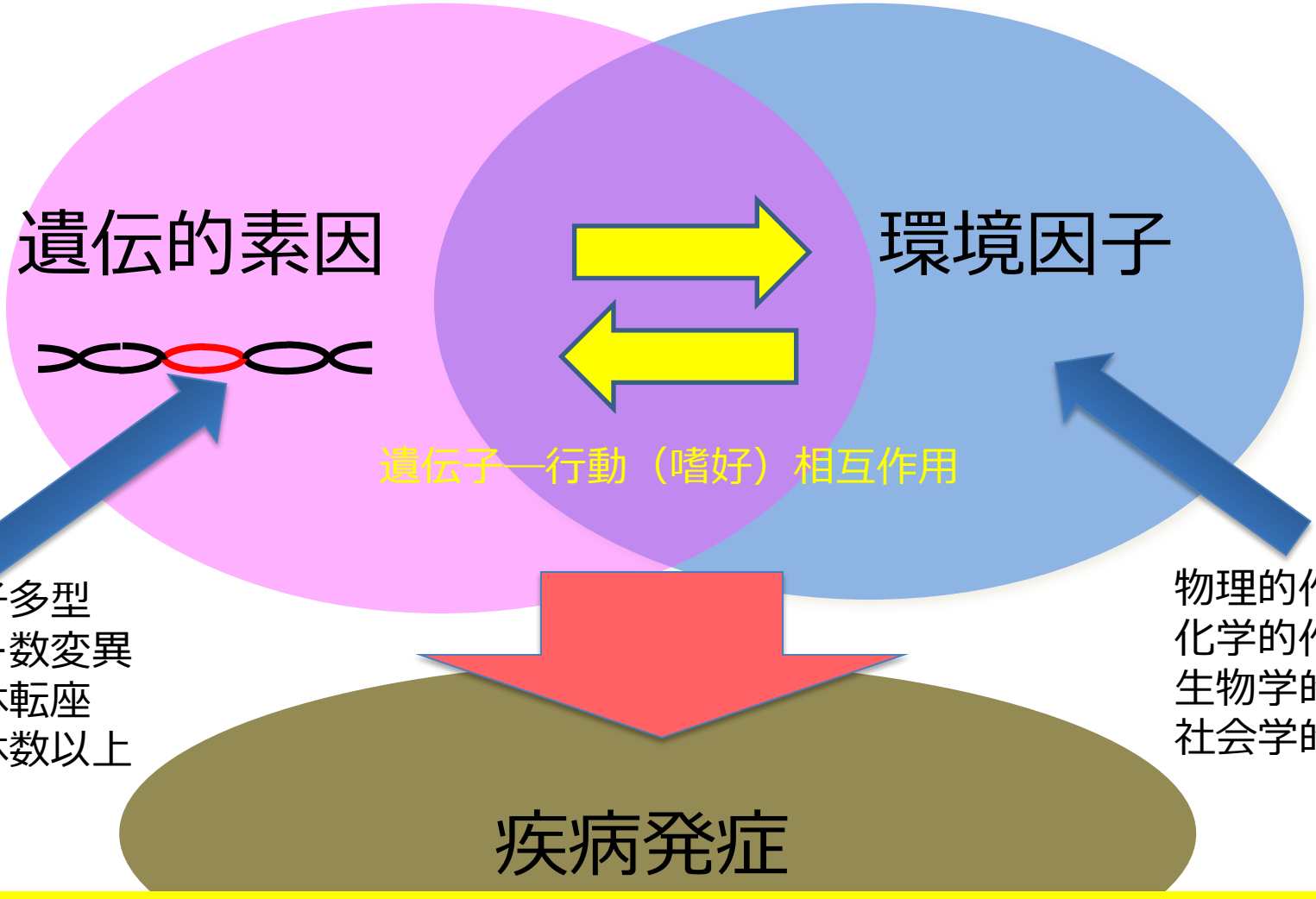
東北メディカルメガバンク(ToMMo)の ゲノム解析の戦略：1000人ゲノム解析

東北メディカル・メガバンク機構
安田　純

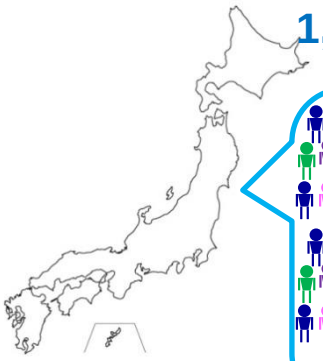
平成27年11月

ヒトの疾病の原因

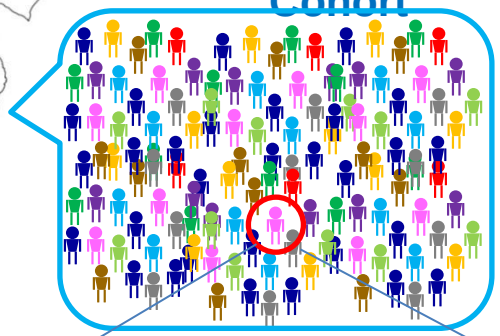


疾病は環境因子と遺伝的素因の両者が原因：どちらも解析対象とすべき

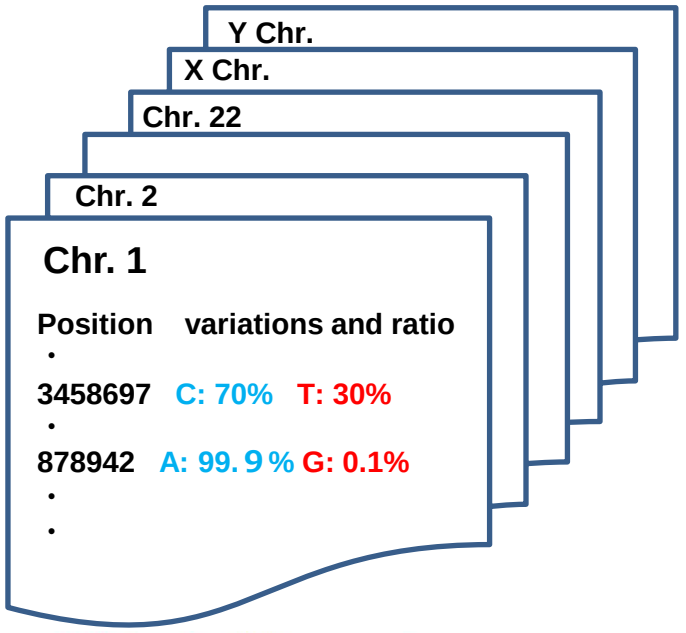
ToMMoのデータ



1,070 from ToMMo regional Cohort



Whole Genome Sequencing

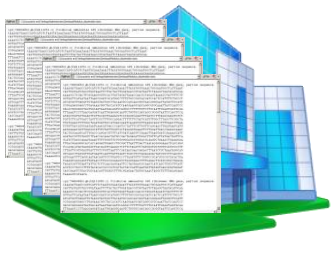


All data will be available at Biobank

Integrative Japanese Genome Variation Database
The genome cohort study of Tohoku Medical Megabank Organization

http://ijgvd.megabank.tohoku.ac.jp/

Biological Specimen: blood, urine, saliva



Electric PHR and physical examination data

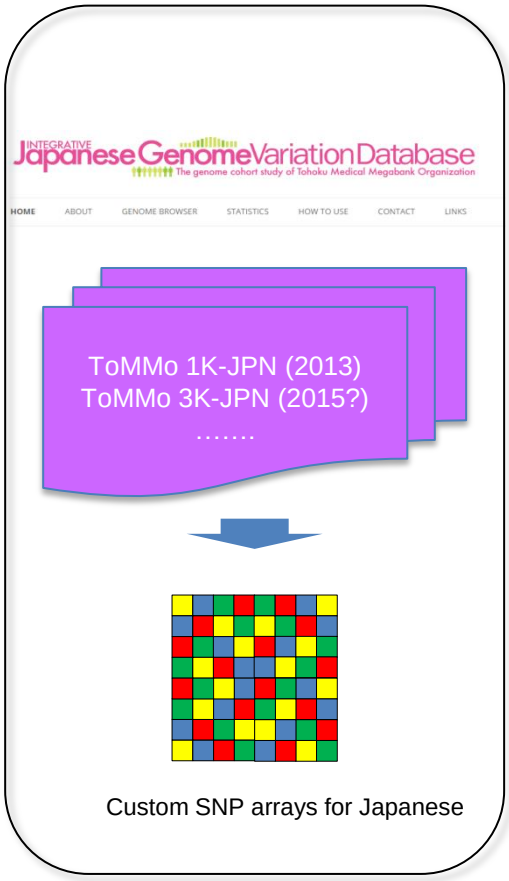
Questionnaire for life styles

[Questionnaires]

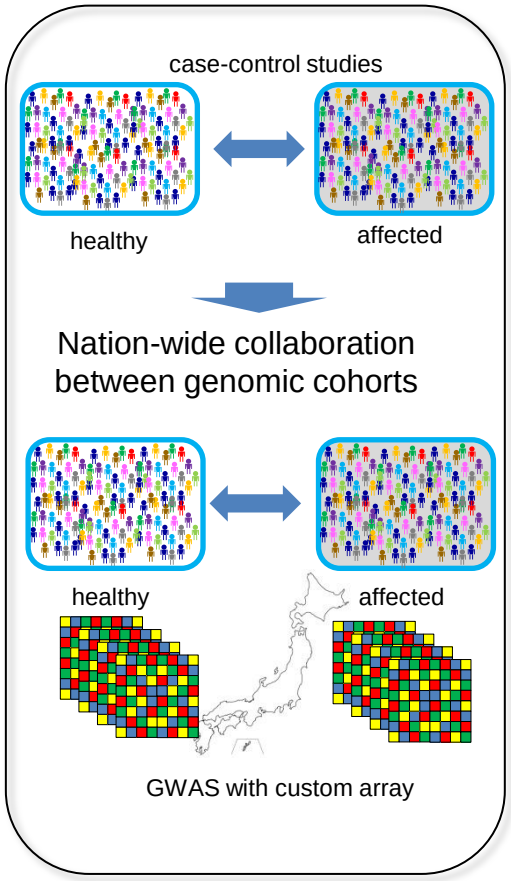
Food (amount and varieties, including preferences)
Daily exercises
Mood (depressive or not)
Quality of Sleep
Tsunami damages

ToMMoのゲノム解析戦略

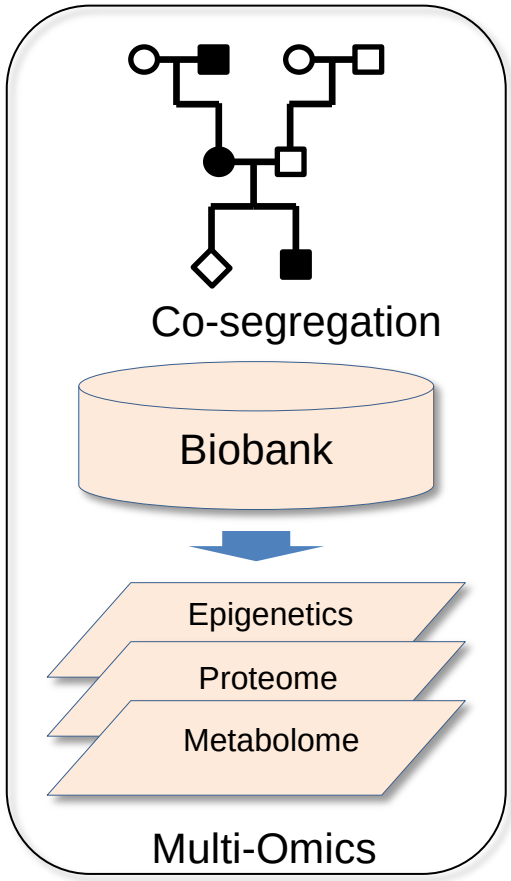
Step 1
Genomic infrastructure



Step 2
ToMMo cohort GWAS



Step 3
Validations



コホート集団ゲノム解析の留意点

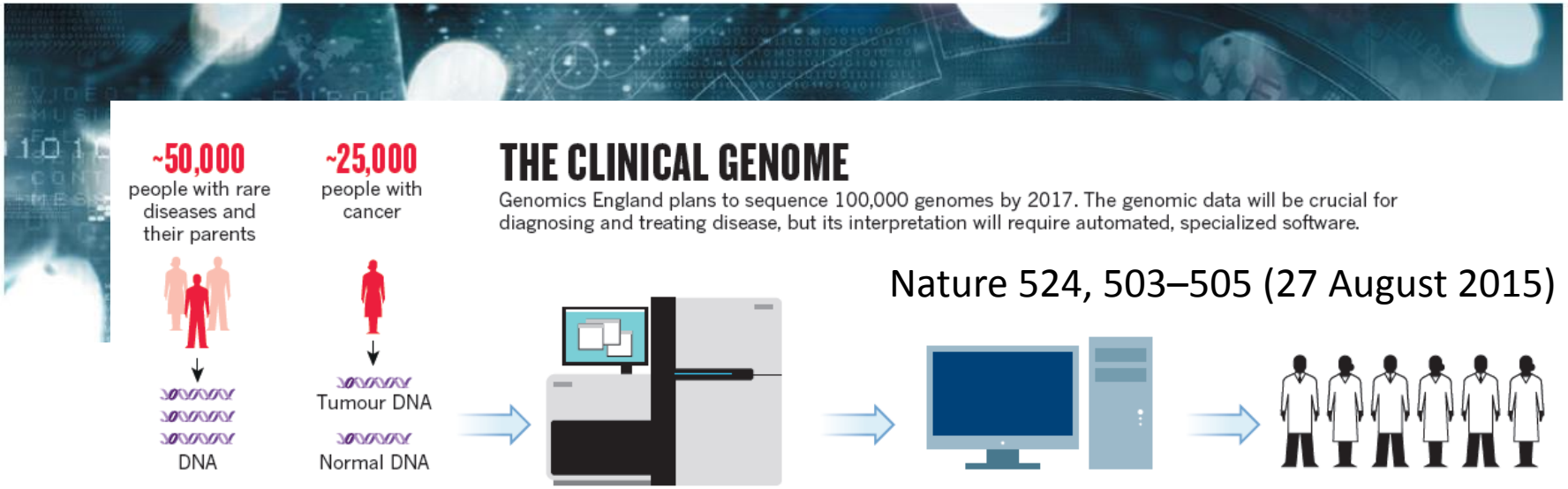
- **参照ゲノムパネルの構築**：ゲノムワイド関連解析の対照として利用
- **血縁関係の排除**：集団内での変異頻度の精密な測定と可能な限り多数のハプロタイプの収集＝近親者は代表者のみとする
- **遺伝学的距離の大きな個人の排除**：階層化の影響をできるだけ除くことで日本人集団ゲノム解析に使いやすい参照パネルとする
- **IDトラッキング**：匿名化された検体（バーコーディング）とHiSeq検体との一貫した突合

TECHNOLOGY FEATURE

THE DNA OF A NATION

The United Kingdom aims to sequence 100,000 human genomes by 2017. But screening them for disease-causing variants will require innovative software.

SERGEY NIVENS/SHUTTERSTOCK



THE CLINICAL GENOME

Genomics England plans to sequence 100,000 genomes by 2017. The genomic data will be crucial for diagnosing and treating disease, but its interpretation will require automated, specialized software.

Nature 524, 503–505 (27 August 2015)

RECRUITMENT OF 75,000 PEOPLE
The 100,000 Genomes Project is recruiting people with cancer and rare diseases. The genomes of both normal and tumour cells will be sequenced in people with cancer.

NEXT-GENERATION SEQUENCING
The Californian company Illumina will use UK-based high-throughput sequencing machines to produce whole-genome sequences and identify genetic variants.

AUTOMATED INTERPRETATION
Four UK and US companies will use specialized software to automatically analyse the genetic variants that may be linked to disease.

CLINICAL INTERPRETATION
Around 2,000 UK scientists and clinicians will pore over the data to validate or better understand how the variants may cause disease before the information is fed back to patients.

Whole genome sequence ($4K/person)

ToMMo 1K-JPN (2013)

ToMMo 3K-JPN (2015)

SNP genotyping with imputation
（Toward $100/person)

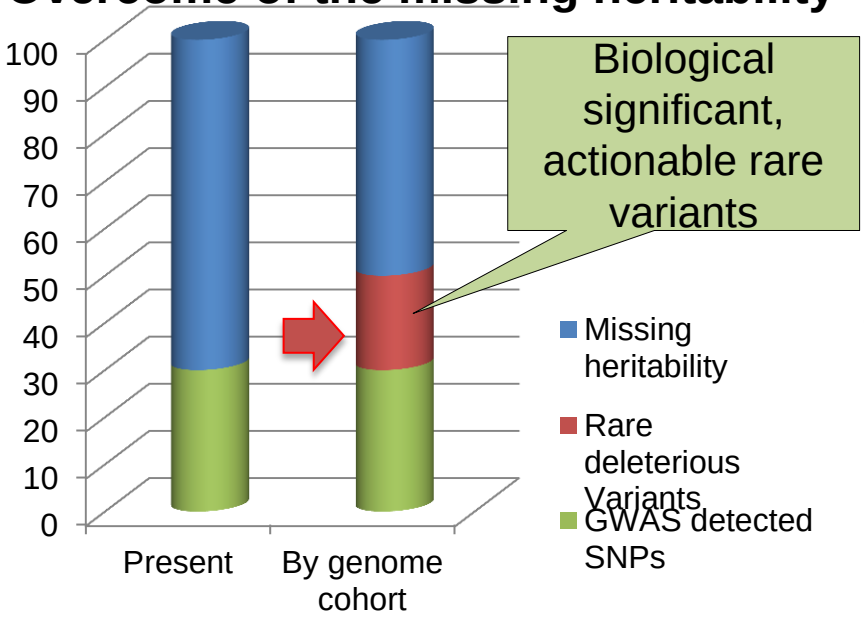
Not detected by SNP arrays

detected

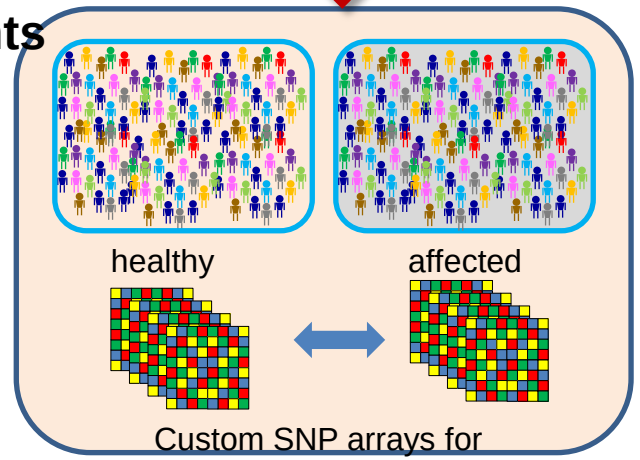
Imputations by LD

Imputed SNP information

Overcome of the missing heritability



Imputation of rare, deleterious variants



GWAS studies

ToMMoのゲノム解析戦略 1

HiSeq2500を使った全ゲノム解析の実際のワークフローの紹介

シークエンス解析のパイプライン

バイオバンクからの検体出庫

SNPアレイ（Omni 2.5-8）での解析

96 dual indexでのライブラリー作成

近親者、非日本人の検出

NGS対象検体の選抜

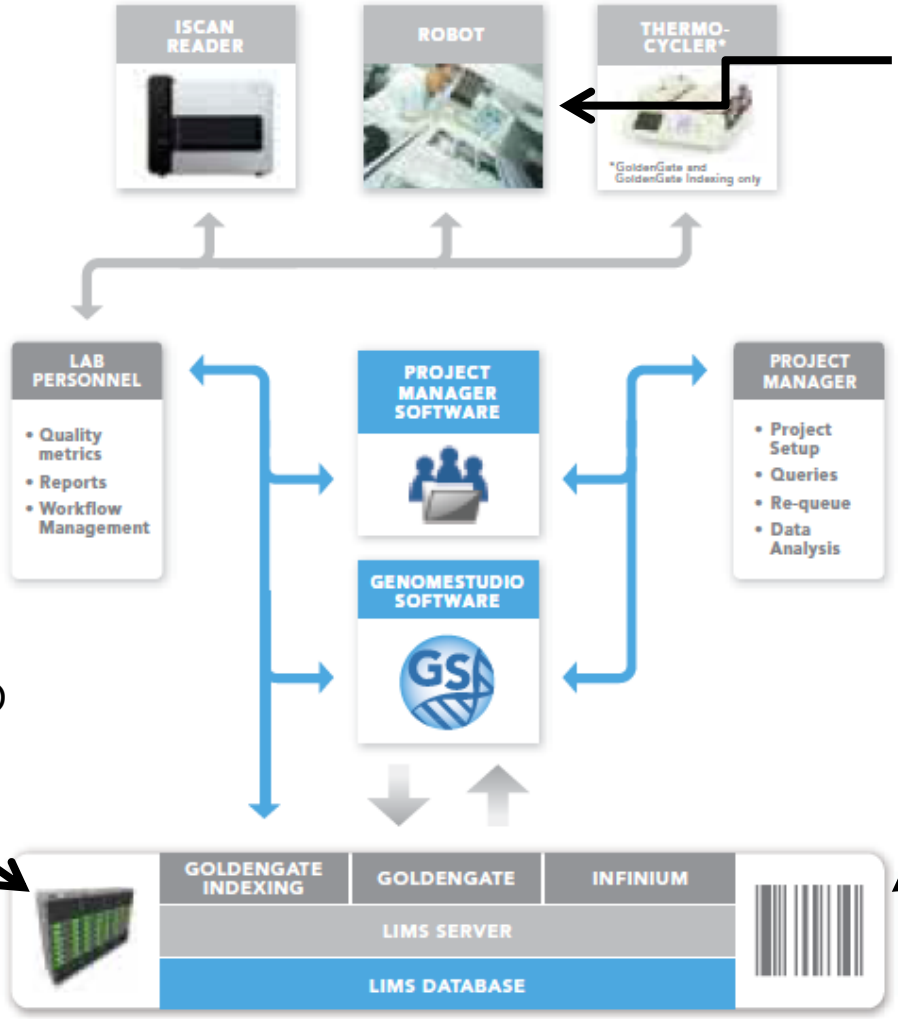
HiSeqでの解析

集団ゲノム解析でのSNPアレイ解析の意義

- **正解セットの提供**：NGSの読み取りの精度確認
- **血縁関係・遺伝学的距離の大きな個人の排除**：PLINKなどのソフトウエアによる近親関係の検出や、類縁集団による構造化の影響をできるだけ除く＝シークエンス試薬の節約
- **近親婚の子孫の排除：**Homozygous stretchの検出によるConsanguinityの検出
- **IDトラッキング**：LIMSによる、匿名化された検体（バーコーディング）と遺伝子型との機械的突合＝HiSeqの検体ロードは手作業であることに留意

イルミナiScanシステムのLIMS

96ウェルプレートから直接サンプルを採取、ビーズチップへのロードまで実施＝検体取り違えの恐れがきわめて低い

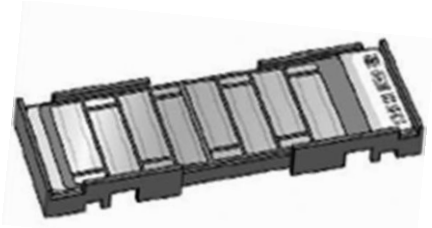


検体のハンドリングはロボットでの自動化

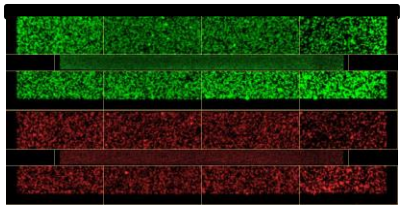
ステップごとの工程管理

検体情報はバーコード管理

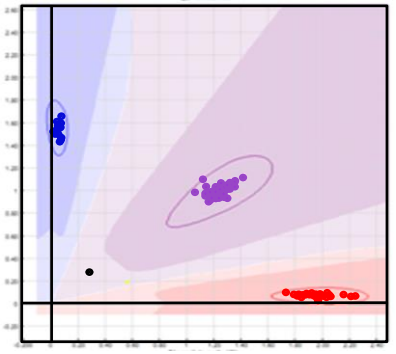
近親および集団から外れた検体の検出手順



ゲノムDNAとマイクロアレイの
ハイブリダイゼーション



ハイブリダイゼーションシグナルの
検出と数値化



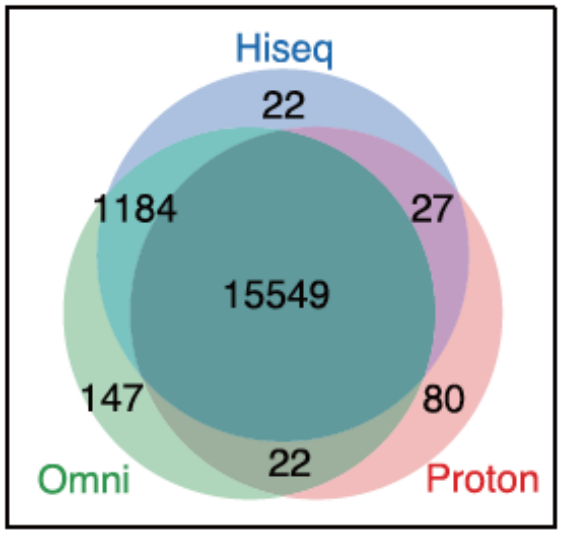
一塩基多型パターンの検出

Omni 2.5-8
ゲノム約250万カ所の
一塩基多型データ

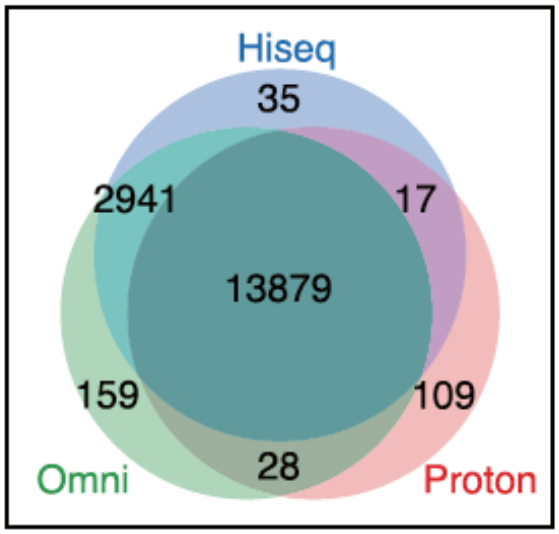
データや検体のクオリティコントロール

- SNP検出時のチェック
- 血縁関係のある個体の排除
- 日本人集団からの外れ個体の排除

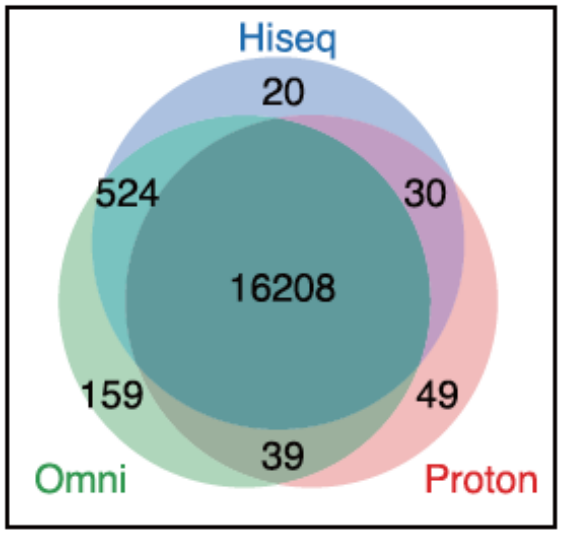
HiSeqとOmni 2.5-8の一致率



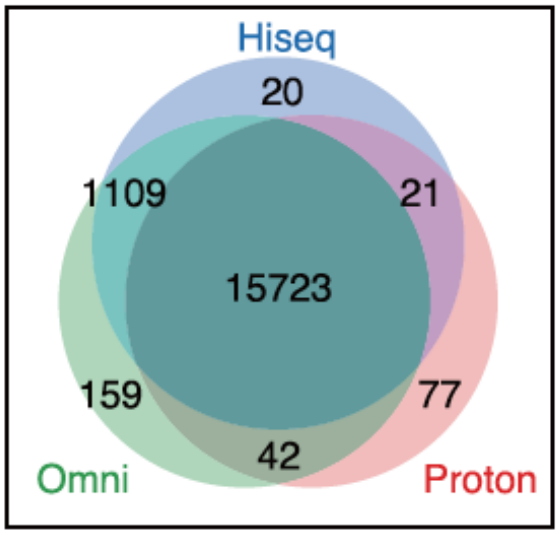
Sample01



Sample02



Sample05

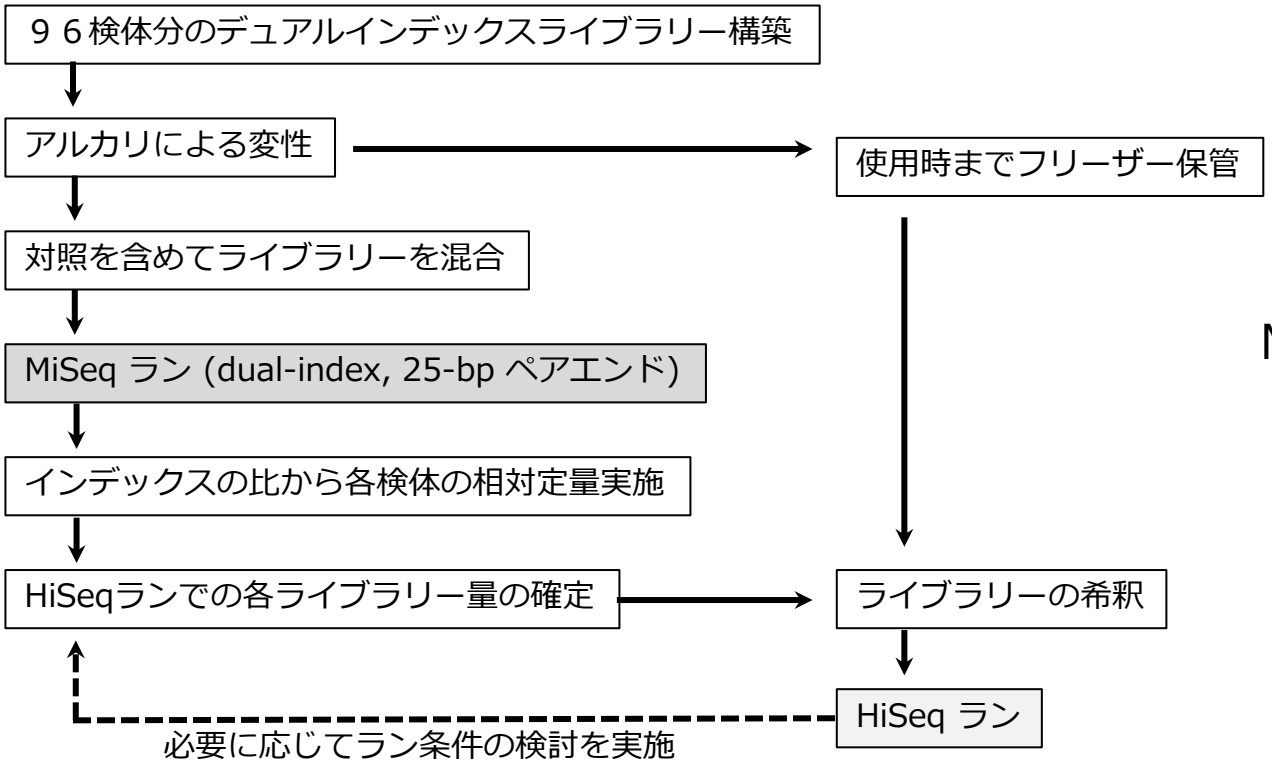


Sample06

- Omni 2.5-8 (250万SNP）での変異コールを正解とするとかなりの部分がProtonでは合わない部分がある。
- HiSeqの情報処理は1000人ゲノムの時よりもかなり簡素なパイプライン（エラーのフィルタリングなし）

Motoike et al., 2014より

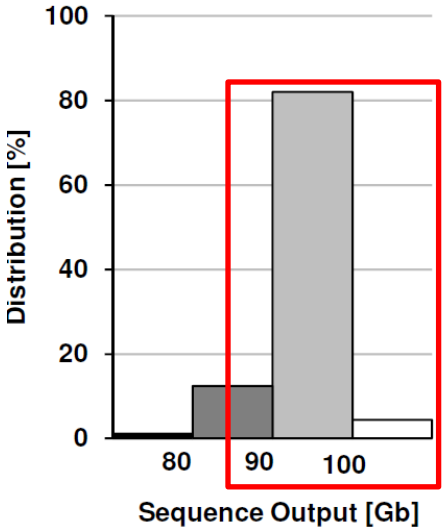
MiSeqによるライブラリQC



適正なクラスター密度のみならず、インサート長も正確に推定可能

(Katsuoka, et al., 2014から改変)

MiSeqでのライブラリーの直接定量



ヒト全ゲノム解読30xを目標＝9割程度達成

ToMMoのゲノム解析戦略2

日本人全ゲノムシーケンス解析結果

ToMMoのゲノム解析情報基盤：1000人ゲノム

nature COMMUNICATIONS

ARTICLE

Received 22 Nov 2014 | Accepted 7 Jul 2015 | Published 21 Aug 2015

DOI: 10.1038/ncomms9018 OPEN

Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals

Masao Nagasaki[1,2,3,*], Jun Yasuda[1,2,*], Fumiki Katsuoka[1,2,*], Naoki Nariai[1,†], Kaname Kojima[1,2], Yosuke Kawai[1,2], Yumi Yamaguchi-Kabata[1,2], Junji Yokozawa[1,2], Inaho Danjoh[1,2], Sakae Saito[1,2], Yukuto Sato[1,2], Takahiro Mimori[1], Kaoru Tsuda[1], Rumiko Saito[1], Xiaoqing Pan[1,†], Satoshi Nishikawa[1], Shin Ito[1], Yoko Kuroki[1,†], Osamu Tanabe[1,2], Nobuo Fuse[1,2], Shinichi Kuriyama[1,2,4], Hideyasu Kiyomoto[1,2], Atsushi Hozawa[1,2], Naoko Minegishi[1,2], James Douglas Engel[5], Kengo Kinoshita[1,3,6], Shigeo Kure[1,2], Nobuo Yaegashi[1,2], ToMMo Japanese Reference Panel Project[#] & Masayuki Yamamoto[1,2]

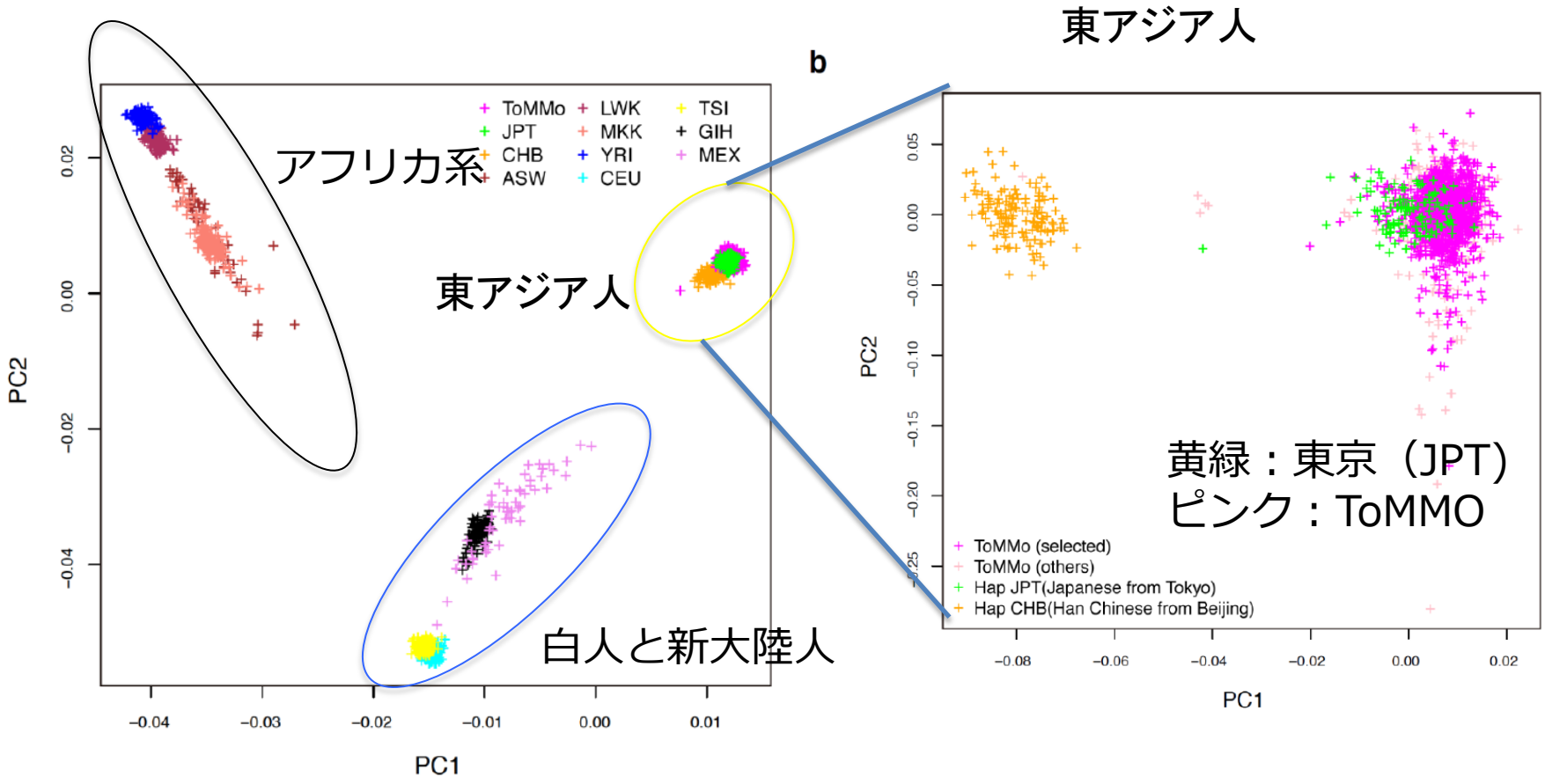
SNPの統計情報（High confidence)

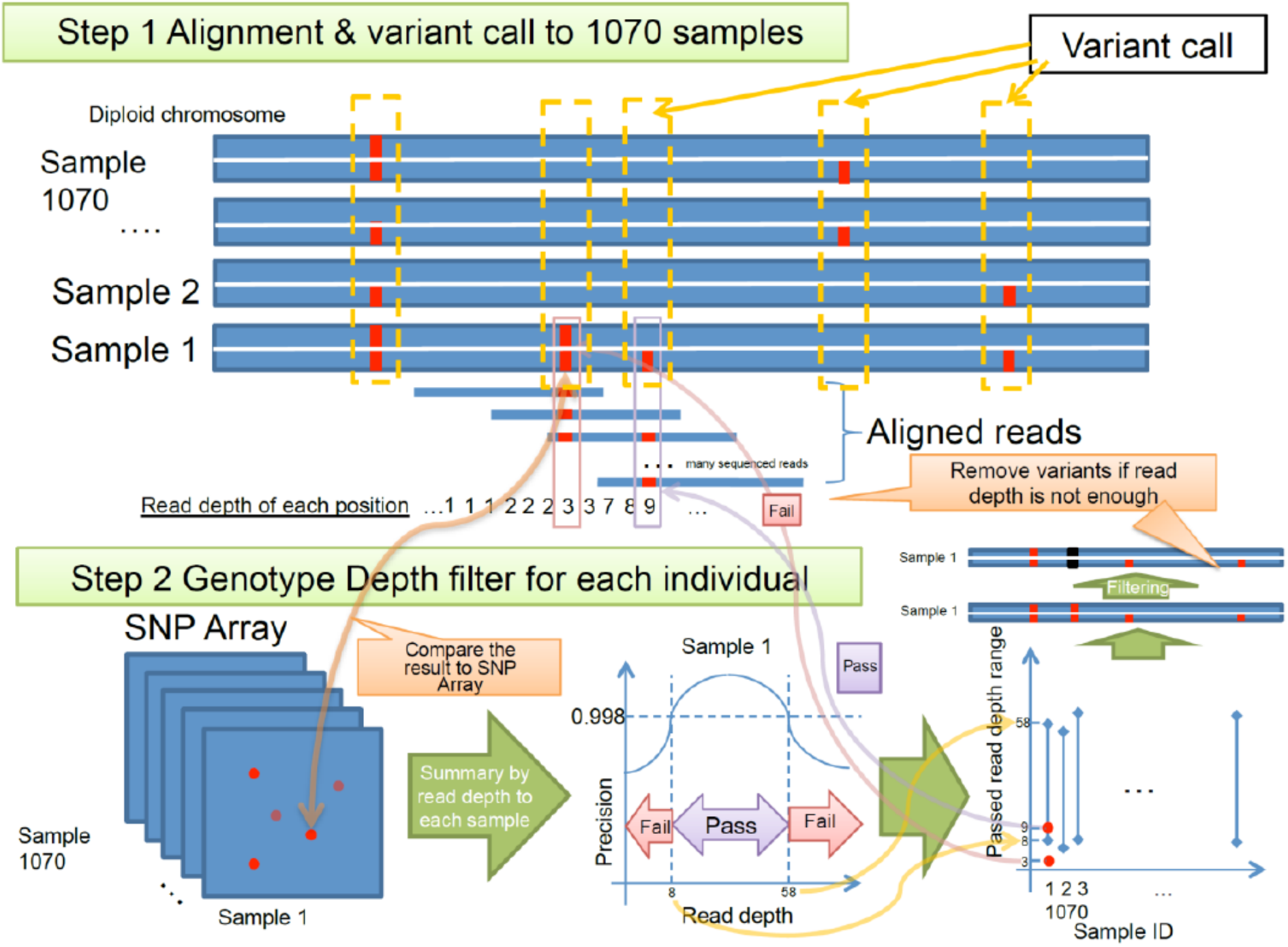
項目	
プロトコール	162 bp PE
インサート長	550 bp
総リード数	100 Trillion

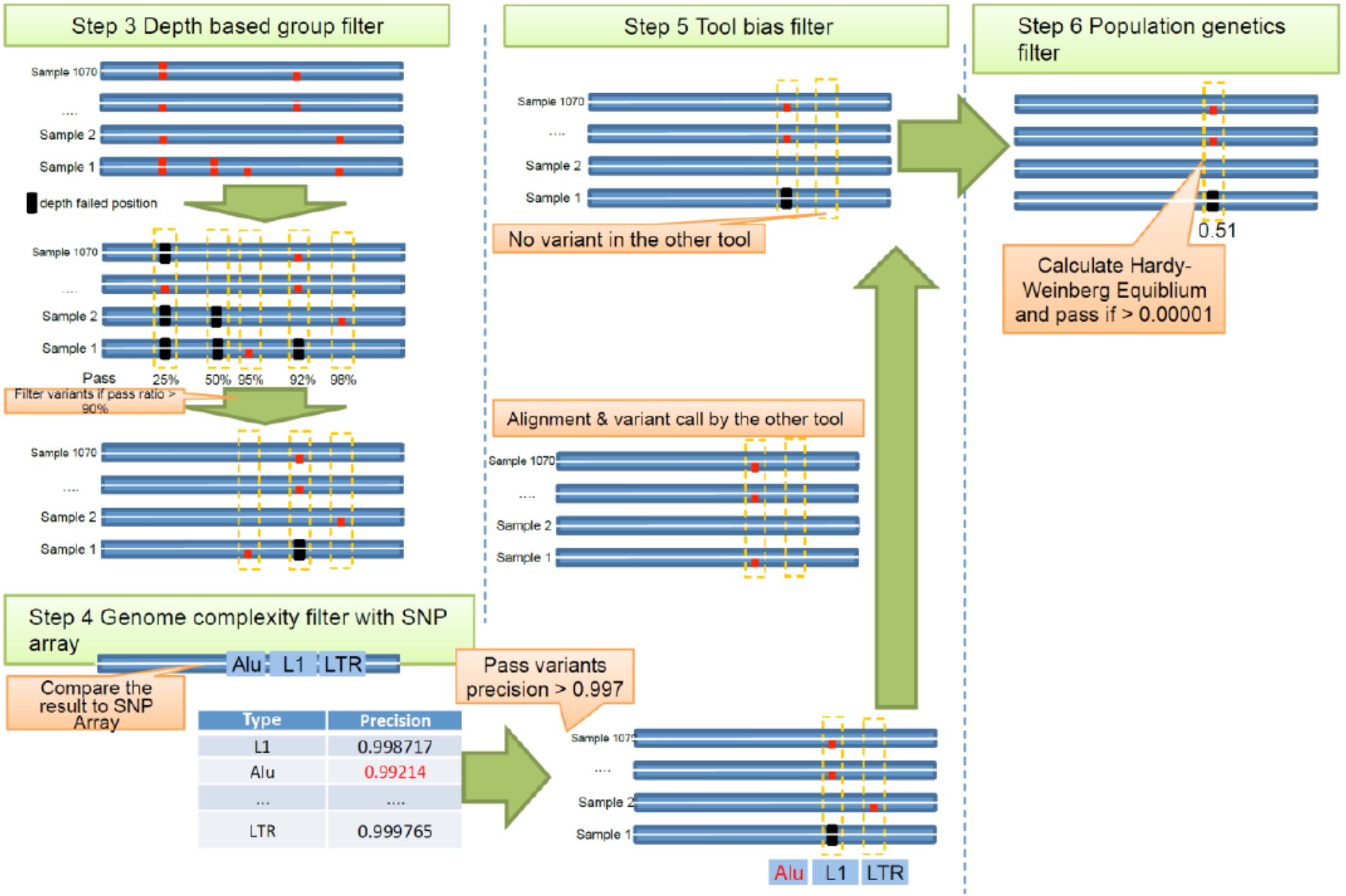
	高信頼SNPセット
平均カバレージ	32.4x
SNPの総数	21,221,195
新規のSNP※	12,001,412
新規のSNPの割合	56.55%
1人あたりのSNP総数	2,716,853
1人あたりのヘテロ接合体数	1,532,773

※dbSNP138との比較

ToMMo1070人の人種上の位置づけ




Step 1 Alignment & variant call to 1070 samples
Variant call
Diploid chromosome
Sample 1070
....
Sample 2
Sample 1
Aligned reads
many sequenced reads
Read depth of each position
...1 1 1 2 2 2 3 3 7 8 9 ...
Fail
Remove variants if read depth is not enough
Step 2 Genotype Depth filter for each individual
Sample 1
Filtering
Sample 1
SNP Array
Compare the result to SNP Array
Sample 1
Pass
0.998
Summary by read depth to each sample
Precision
Fail
Pass
Fail
8
58
Read depth
Passed read depth range
58
9
8
3
1 2 3
...
1070
Sample ID
Sample 1
Sample 1070

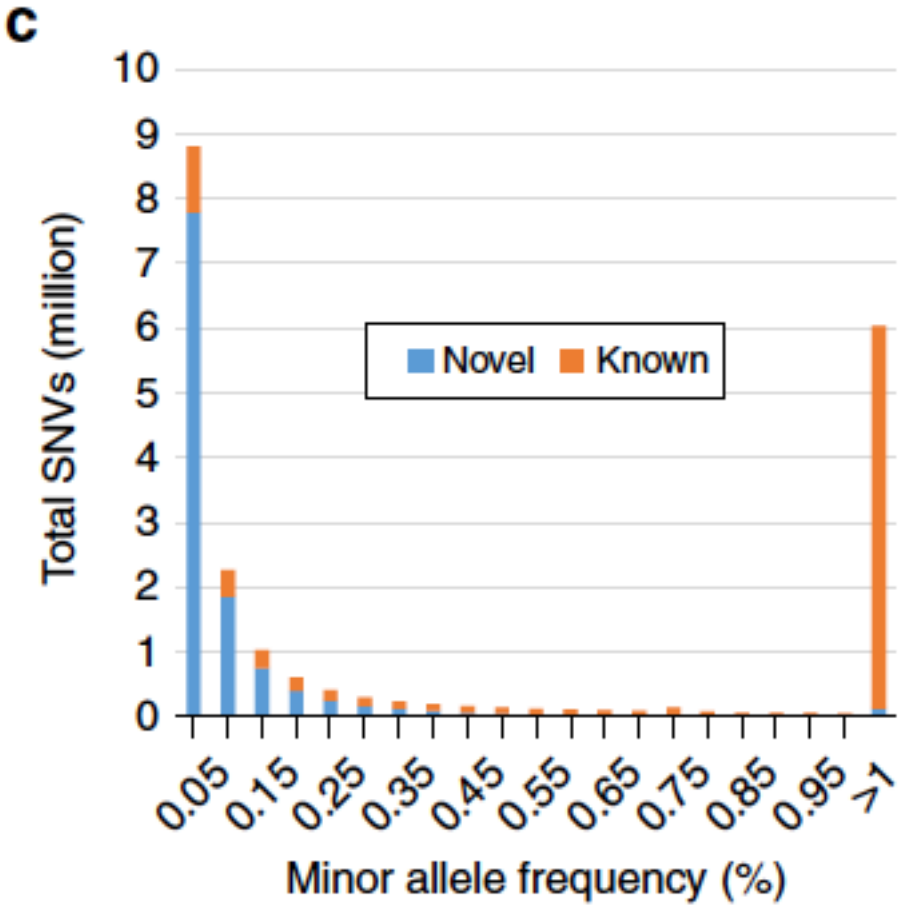

Step 3 Depth based group filter
Sample 1070
....
Sample 2
Sample 1
depth failed position
Pass
25%
50%
95%
92%
98%
Filter variants if pass ratio > 90%
Step 4 Genome complexity filter with SNP array
Alu
L1
LTR
Compare the result to SNP Array
Type
Precision
L1
0.998717
Alu
0.99214
...
....
LTR
0.999765
Pass variants precision > 0.997
Step 5 Tool bias filter
No variant in the other tool
Alignment & variant call by the other tool
Step 6 Population genetics filter
0.51
Calculate Hardy-Weinberg Equiblium and pass if > 0.00001

反復配列へのフィルタリング効果

Filter detail	Total	Known	Novel	Novelty rate	Pass SNVs
Raw call with bowtie2 + bcftools	27,490,104	11,914,146	15,575,958	56.66%	100.00%
Filter variants in unreliable depth of coverage in the sample. (FDR<0.2%)	26,939,185	11,824,964	15,114,221	56.10%	98.00%
Filter variants in unreliable depth of coverage in population.	25,568,721	11,194,027	14,374,694	56.22%	93.01%
Filter variants categorized into low precision genomic region. (FDR <0.3%)	21,660,722	9,509,974	12,150,748	56.10%	78.79%
Intersect variants with other variant caller.	21,504,896	9,483,893	12,021,003	55.90%	78.23%
Remove SNVs with HWE < 0.00001	21,221,195	9,219,783	12,001,412	56.55%	77.20%

主として反復配列中に存在するSNPを排除＝全体の14%程度が取り除かれた

新規SNVと頻度



新規のSNPはシングルトンなど希少なものががほとんど

配列機能と希少変異の割合

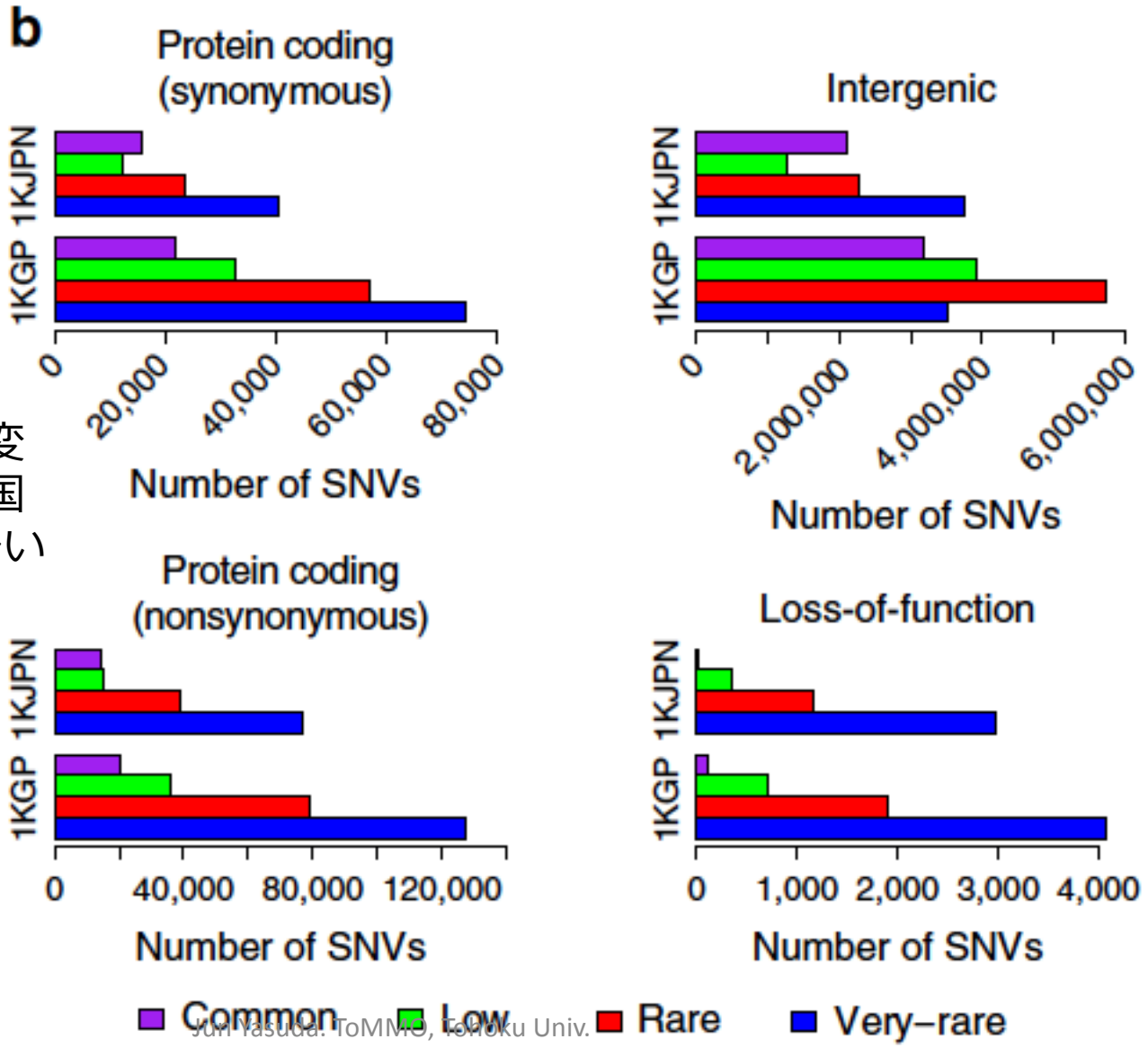
ToMMoの結果

国際1000人の結果

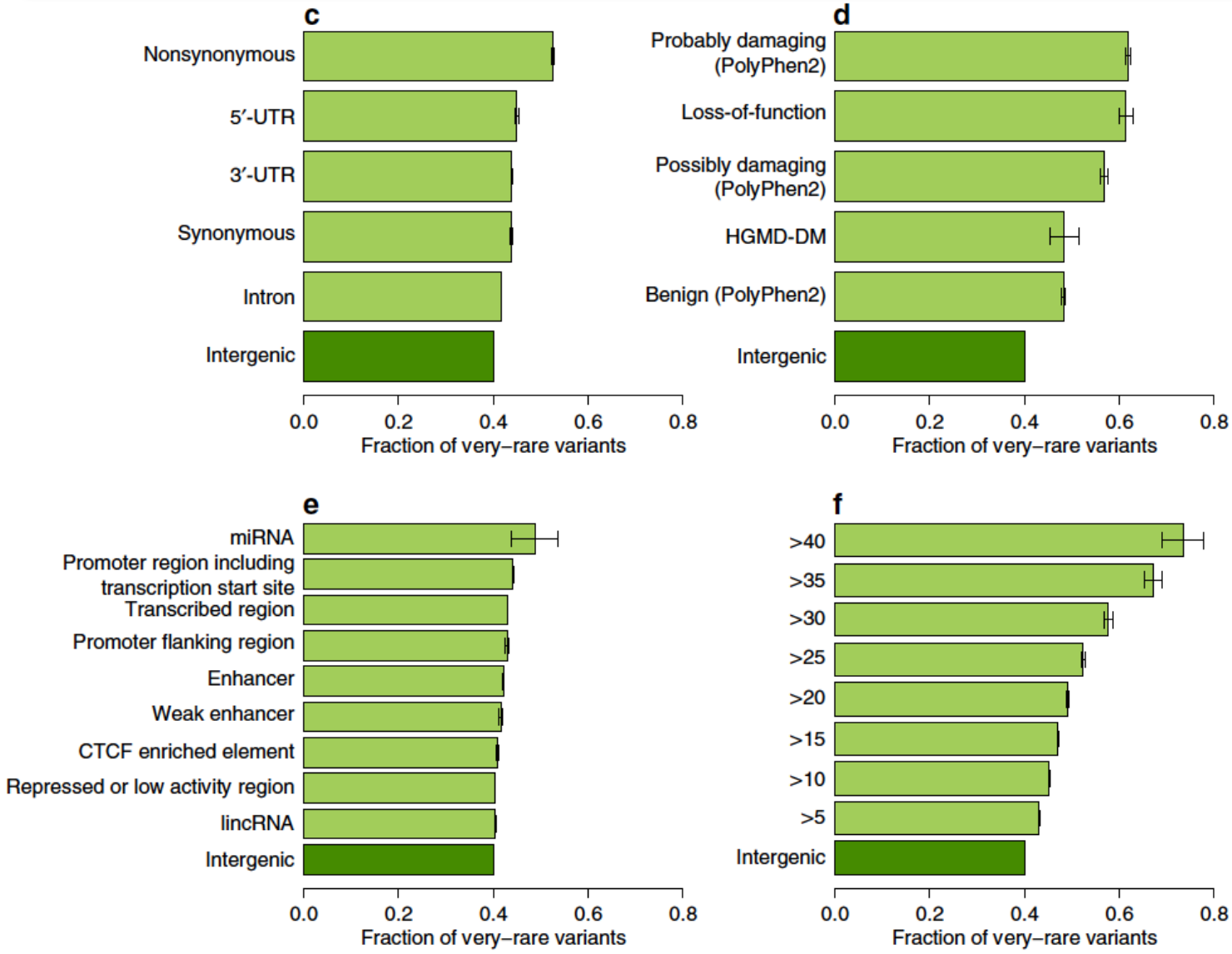
人種によって固有の変異があるため全体に国際1000人のほうが多い

ToMMoの結果

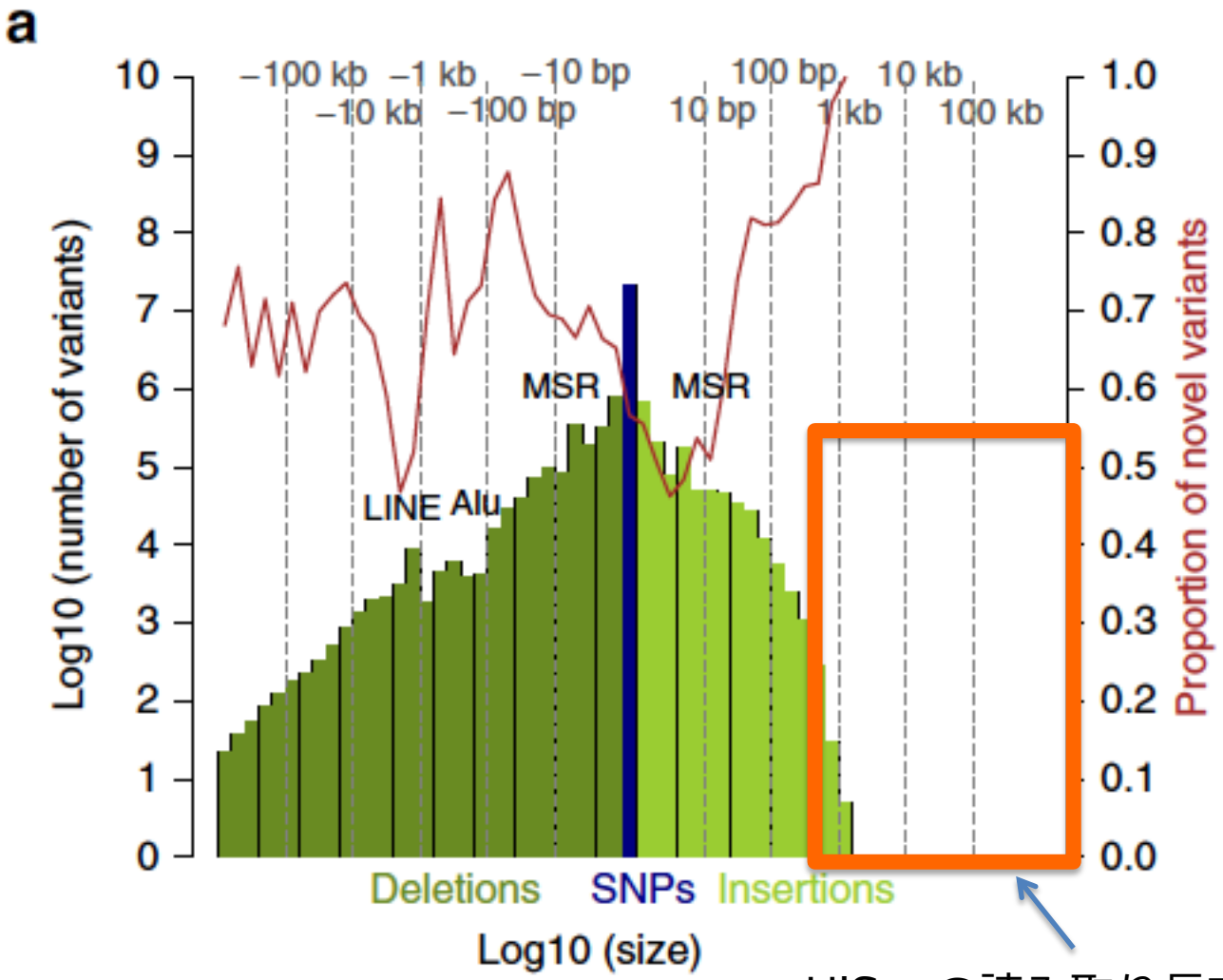
国際1000人の結果



希少変異の目立つ属性


c
Nonsynonymous
5′-UTR
3′-UTR
Synonymous
Intron
Intergenic
0.0
0.2
0.4
0.6
0.8
Fraction of very−rare variants
d
Probably damaging (PolyPhen2)
Loss-of-function
Possibly damaging (PolyPhen2)
HGMD-DM
Benign (PolyPhen2)
Intergenic
0.0
0.2
0.4
0.6
0.8
Fraction of very−rare variants
e
miRNA
Promoter region including transcription start site
Transcribed region
Promoter flanking region
Enhancer
Weak enhancer
CTCF enriched element
Repressed or low activity region
lincRNA
Intergenic
0.0
0.2
0.4
0.6
0.8
Fraction of very−rare variants
f
>40
>35
>30
>25
>20
>15
>10
>5
Intergenic
0.0
0.2
0.4
0.6
0.8
Fraction of very−rare variants

挿入欠失配列の頻度分布



HiSeqの読み取り長では検出不能

ToMMoのゲノム解析戦略3

日本人全ゲノム解析の意義

日本人一人当たりの病的変異数

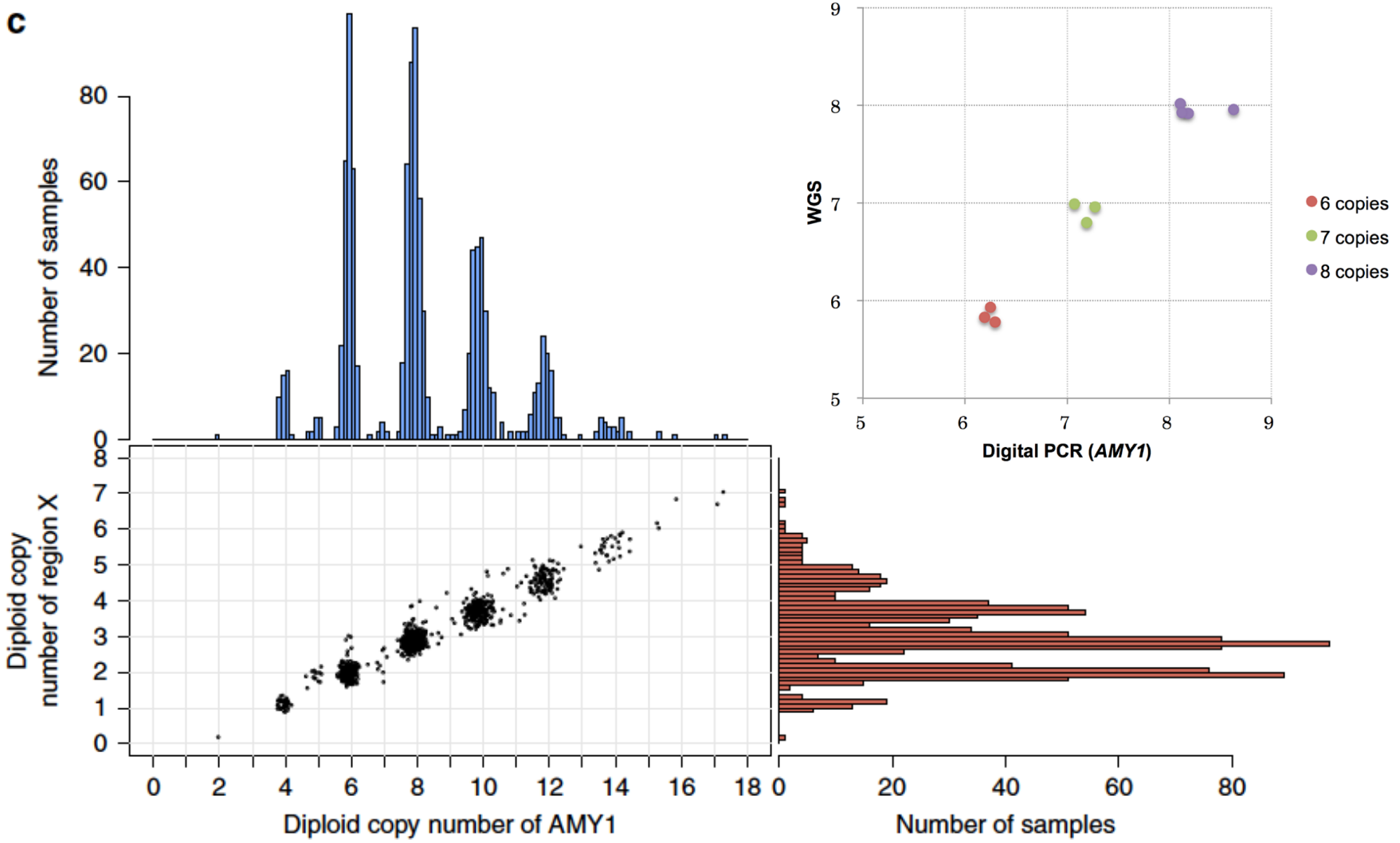
頻度	<0.1%	0.1~0.5 %	0.5~5%	>5%
停止発生	2.39	1.56	8.68	29.0
疾病関連変異	0.64	1.04	4.76	3.18

疾病に関連する変異は、一人平均10-20個近く持っている＝昔から推定されていた頻度とそう変わらない

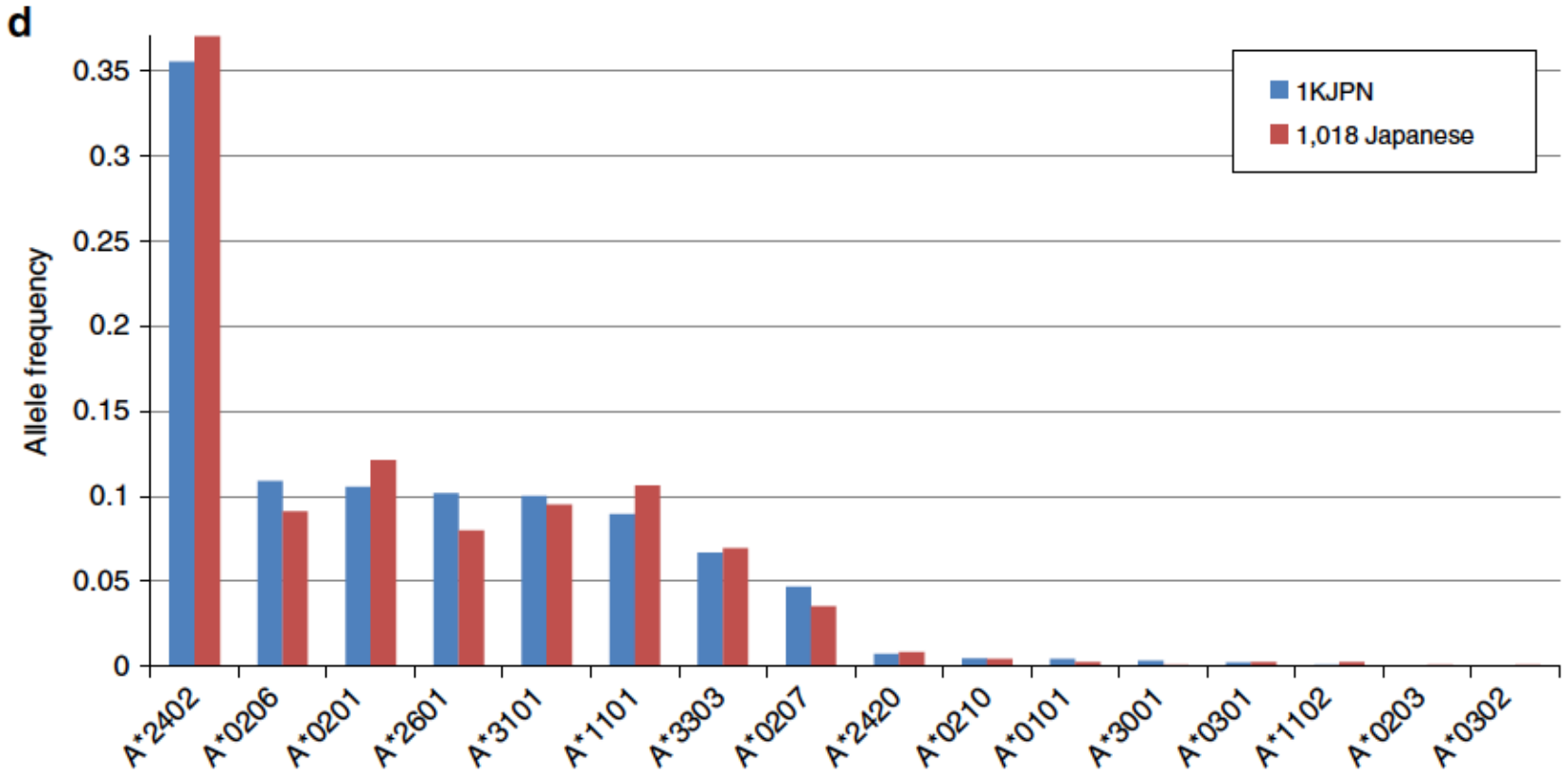
注意：これらのうち、本当に疾病にかかわっているものは一部のみ

Nagasaki et al., 2015より改変

読み取り深度を活用したCNVの検出例

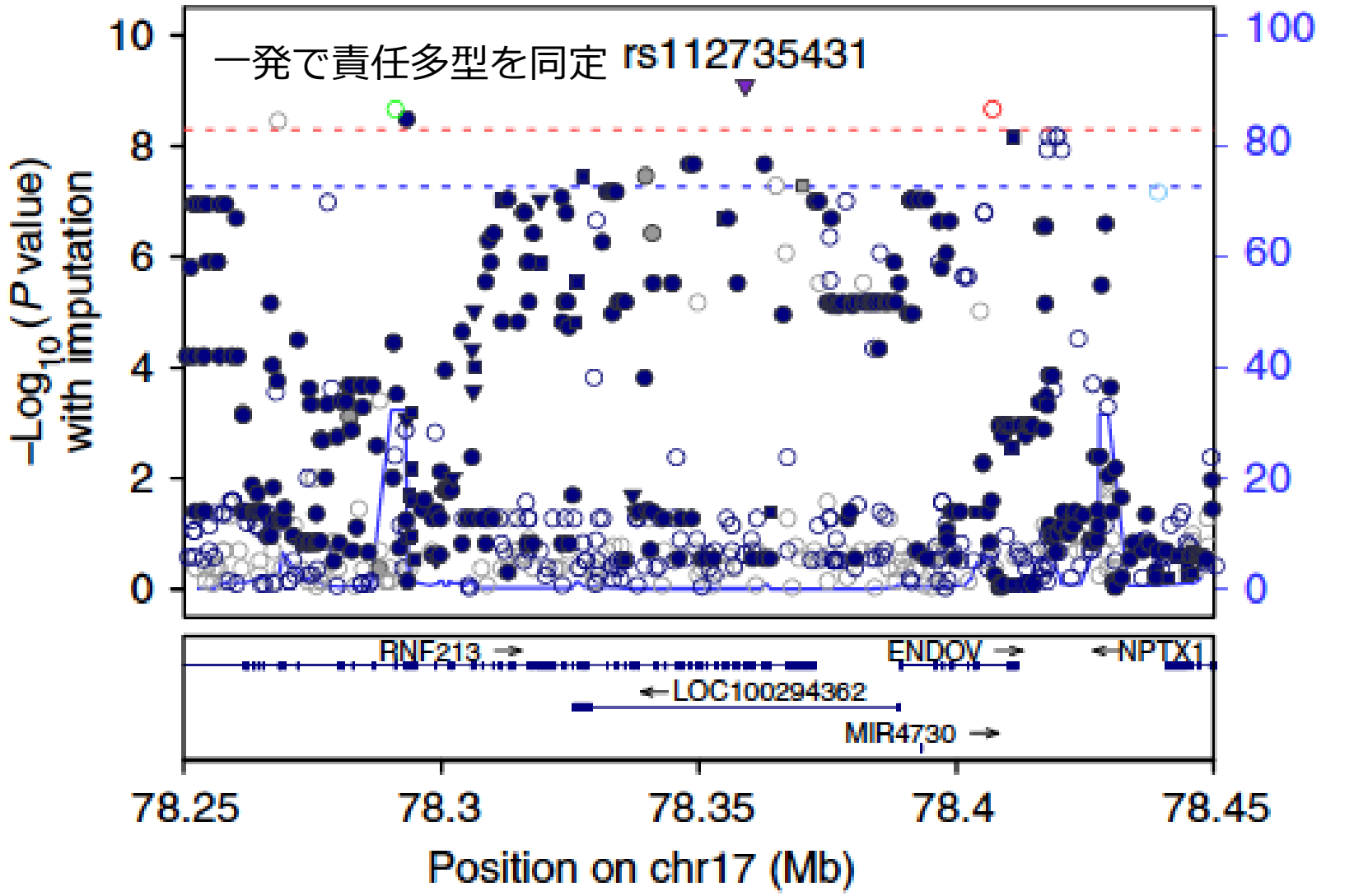


HiSeqデータでのHLA-A遺伝子座の解析



過去の日本人のデータと整合性がある（既知のハプロタイプのみ）

インピュテーションの効果



まとめ

- 大集団の全ゲノム解読にはSNPアレイ解析での集団構造の分析や近親排除が必要。LIMSによる検体のIDトラッキングがきわめて有用
- HiSeqライブラリーQCには、MiSeqでの定量が有力
- 日本人1070人の全ゲノム解読を実施、総計2100万か所の一塩基置換を同定し、希少変異に有害変異が目立つことを見出した
- 日本人など、農耕民族で増加しているAMY1遺伝子のコピー数についてNGSの深度情報から正確に推定可能
- HLAクラスI遺伝子のハプロタイプもNGSのリードから正確に確認できた
- もやもや病の関連遺伝子変異をimputationによって元データから推定することができた
- 今後、日本人に特徴的な有害変異などの探索を進める予定

謝辞

シークエンス解析室
安田純、勝岡史城、檀上稲穂、黒木陽子、斎藤さかえ、横澤潤二、齋藤るみ子、津田薫
インシリコ解析室
長崎正朗、小島要、河合洋介、成相直樹、佐藤行人、三森隆広、山口由美、柴田朋子
オミックス解析室
田邊修、小柴生造、三枝大輔、加藤泰丈
スーパーコンピュータ運営室
木下賢吾、元池育子、城田松行

東北大学東北メディカル・メガバンク機構
山本雅之、八重樫伸生、呉繁夫、布施昇男、峯岸直子、荻島創一、高井貴子、寶澤篤、
栗山進一、清元秀泰、菅原準一、瀧靖之、坪井明夫、鈴木洋一、川目裕、田宮元

他、多数の皆様のおかげです。