The GenomeAsia 100K Project enables genetic discoveries across Asia

The underrepresentation of non-European individuals in human genetic studies limits the applicability of the results for a large proportion of the world's population. Reference genome datasets are needed to characterize population-specific variation, enable efficient imputation of variants that are not directly genotyped, and extend genome-wide association studies (GWAS) to additional populations. The value of population-specific reference datasets is well recognized and projects based in the United States and Europe have provided deep characterization of specific populations (for example, Ashkenazi Jews and individuals from the Netherlands and Iceland) and, in particular, data from individuals of Nordic countries have provided examples of how reference genome datasets can be used to drive comprehensive genetic studies across an entire population. In Africa, populations show complex genetic patterns, smaller blocks of linkage disequilibrium and higher levels of heterozygosity, which provides unique value for genetic studies. Across the continent, early reference genome datasets for diverse populations are being built as part of H3Africa and other studies. A Korean reference genome as well as Japanese and Chinese reference genome datasets have been created, and the formation of large biobanks such as BioBank Japan and the China Kadoorie Biobank will accelerate the pace of discovery of disease associations across east Asia.

A shared recognition of the value of coordinated efforts and the need for reference genome datasets that would be useful for the complex populations of Asia has led to the formation of the GenomeAsia consortium (http://www.genomeasia100k.com). The consortium serves to facilitate and coordinate sequencing efforts among consortium members to maximize the value of the genomic sequence data that is produced and to facilitate efforts by national or other regional groups. Here we describe the GenomeAsia Pilot (GAsP) project, which consists of analyses of the whole-genome sequencing data of 1,739 individuals from 219 population groups across Asia, with the ultimate goal of providing a useful genomic resource and facilitating genetic studies in Asia. We use the data that was generated in this pilot to analyse population structure and history, and as the basis for designing larger-scale genomic studies. Furthermore, we explore disease-associated loci as an initial comparison of differences between populations. We show that the variant data produced by this project improve variant filtering for the discovery of disease-associated genes of rare diseases. We show that Asia has sizable founder populations and that further studies in these populations may be useful for the discovery of rare-disease-associated genes. We also report an initial survey of loss-of-function alleles found in the GAsP project.

The GAsP dataset

For the GAsP project, we generated 1,267 high-coverage (average 36×) whole-genome sequences and analysed these together with 596 publicly available human genome sequences from previous sequencing studies (Supplementary Information 1, 2 and Supplementary Tables 1a–c, 2a). The 1,739 samples were enriched for individuals from population isolates to capture the broadest wealth of genetic diversity; the dataset includes 598 sequences from India, 156 from Malaysia, 152 from South Korea, 113 from Pakistan, 100 from Mongolia, 70 from China, 70 from Papua New Guinea, 68 from Indonesia, 52 from the Philippines, 35 from Japan and 32 from Russia (Fig. 1a–c and Supplementary Table 1a–c). To facilitate comprehensive and comparative analysis of human genetic variation, we included sequencing data from African, European and American samples (Supplementary Table 1a, b). The sequenced samples originate from 7 global regions, 64 different countries of origin and 219 population groups. About 80% of the samples come from Asia and emphasize population groups that are underrepresented in previous genetic studies (Fig. 1a, b, Supplementary Tables 1a–c, 2b and Supplementary Information 1, 2). Each global region and population group was assigned a unique three-letter code for future reference (see Supplementary Table 1a for three-letter code designations). Within Asia, the sampling of many distinct population groups allowed us to analyse the relationship between geography, physical characteristics and genetic variation. In south and southeast Asia, in particular, we sampled across diverse populations to gather new insights into how groupings defined on the basis of caste and language relate to genetic diversity, admixture with extinct hominins and other genetically described characteristics.
what was found in a previous study that was based on single-nucleotide polymorphism arrays (SNP) arrays\(^5\). The high levels of Denisovan ancestry in Melanesians and the Aeta are consistent with an admixture event that is ancestral to both\(^23\); however, two lines of evidence suggest that the ancestors of the Aeta experienced a second Denisovan admixture event. First, multiple analyses found that the Aeta are genetically more similar to populations without appreciable Denisovan ancestry (for example, Igorot, Malay and Malay Negrito groups) than they are to Melanesians (Supplementary Information 3, 6). This can be explained by more recent gene flow from other populations without Denisovan ancestry. However, such gene flow would reduce the levels of Denisovan admixture below that found in Melanesians. More directly, we find that putative Denisovan haplotypes that are unique to the Aeta (n = 962) are significantly longer than putative Denisovan haplotypes shared between Aeta and Papuans (n = 596, mean = 16.1 kb compared with mean = 14.1 kb, Mann–Whitney U-test, P < 10\(^{-9}\)). This can be explained by more recent gene flow from other populations without Denisovan ancestry. However, such gene flow would reduce the levels of Denisovan admixture below that found in Melanesians. More directly, we find that putative Denisovan haplotypes that are unique to the Aeta (n = 962) are significantly longer than putative Denisovan haplotypes shared between Aeta and Papuans (n = 596, mean = 16.1 kb compared with mean = 14.1 kb, Mann–Whitney U-test, P = 596, mean = 16.1 kb compared with mean = 14.9 kb, Mann–Whitney U-test, P = 9.96 × 10\(^{-3}\}) (Supplementary Information 10), supporting a scenario in which a second admixture event between the Aeta and Denisovans happened after the separation of the Aeta and Melanesians. Two distinct Denisovan admixture events are most consistent with Homo sapiens Denisovans interacting within southeast Asia\(^23\), making it likely that admixture occurred within Sundaland (Fig. 2b) or even farther east\(^24,25\).

A recent study found a slightly increased amount of Denisovan ancestry in south Asians compared with a priori expectations\(^26\). We examined whether this was correlated with either language or social and/or caste status. South Asian samples were grouped into individuals who speak Indo-European languages and individuals who speak non-Indo-European languages (excluding individuals who speak Tibeto-Burman languages), as well as four social or cultural groups: tribal (Adivasi) groups, lower-caste groups, high-caste groups and Pakistani groups (Indo-European language speaking only). We found that the average levels of Denisovan ancestry were significantly different between the four social or cultural groups (Mann–Whitney U-test, P < 10\(^{-8}\) for all pairwise comparisons; Fig. 2c and Supplementary Information 10). Our results are consistent with the scenario that Indo-European-speaking migrants who entered the subcontinent from the northwest admixed with an indigenous South Asian (ancestral south Indian)\(^27,28\) group who had higher levels of Denisovan ancestry.

**Medical relevance**

We evaluated the use of GASp dataset in disease-associated genetic studies and medically relevant applications to determine how the results of larger continuing GenomeAsia studies can be used to improve human health (Supplementary Table 4a). We annotated high-quality variants using public databases including ExAC (Exome Aggregation Consortium)\(^29\), gnomAD\(^30\), 1000 Genomes Project\(^3\), ESP (NHGRI GO Exome Sequencing Project)\(^30\) and dbSNP (Extended Data Fig. 2) and focused...
on coding-sequence variants. Overall 23% of protein-altering variants in GAsP were not found in these data sources. As expected the majority of coding variants were singletons or very rare (Extended Data Fig. 2). However, the absolute numbers of novel variants with a minor allele frequency (MAF) > 0.1% within our pan-Asian dataset is large (n = 194,585), and these are frequent enough to be of relevance for large-scale genetic association studies. We also searched for variants present at low frequency in the overall dataset that are present at significantly higher allele frequencies in one or more of the population groups. We found an additional 144,329 novel variants with MAF > 1% in the full GAsP dataset that were present at a frequency of greater than 1% within populations grouped by geography; South Asia, Southeast Asia, Northeast Asia or Oceania (see Supplementary Table 1a for description of samples and population groups included in each geographically defined set). These geographical regions contain many diverse population groups, and additional studies are needed to characterize patterns of genetic variation in these groups and disease relevance.

In rare disease genetics, databases are used to filter based on allele frequency with the idea that common alleles are unlikely to be responsible for rare highly penetrant disorders; however, in the absence of appropriate population reference datasets, allele frequencies can be misclassified and may lead to false disease associations. We explored whether the GAsP variant dataset can improve the ability to identify disease-relevant variants in Asian cohorts. We analysed 152 exomes from individuals participating in the Indian Maturity Onset Diabetes in the Young (MODY) project. When both the gnomAD and GAsP datasets were used for filtering (MAF > 0.1%), we reduced the set of remaining candidate variants by approximately twofold in comparison to using the gnomAD dataset alone (Fig. 3a). In this process, we identified a common population polymorphism in NEURO1 (H241Q) that is probably benign but that was previously reported to be medically relevant. We annotated variants that were identified in the GAsP dataset against the Human Gene Mutation Database (HGMD) disease-causing pathological and ClinVar pathogenic variants. This analysis identified 732 variants (686 SNPs and 46 insertions or deletions (indels)) in 514 genes (Fig. 3b, Supplementary Table 4b, c and Supplementary Information 11). We compared the 732 pathogenic variants against the gnomAD, ExAC, 1000Genomes, ESP, dbSNP, ALSPAC, TwinsUK and 1000Japanese databases to remove variants that occurred at <1%, focused on those with allele frequencies >0.15% in GAsP (38 variants), and reviewed them against the criteria defined by the American College of Medical Genetics (ACMG). This resulted in reclassification of 11 of the 38 variants (Supplementary Table 4d). We examined the geographical distribution of the remaining, revalidated but high-frequency, pathogenic disease-associated variants. As expected, most of these variants were highly enriched in Asia. For example, an HBB variant (chromosome 11: 5248155 c.92+5G>C) associated with β-thalassaemia is found almost exclusively in south Asians and at a lower frequency in southeast Asians (Fig. 3c).

We also examined our dataset for novel variants in genes known to be associated with cancer risk. We found 13 unique variants in 6 genes from 17 samples. This included frameshift, stop-gained and essential splice-site mutations in BRCA2 (n = 9), BRCA1 (n = 1), ATM (n = 2), BLM (n = 1), NBN (n = 2) and PMS2 (n = 2) (Fig. 3d and Supplementary Table 4e). Of the two PMS2 essential splice variants, one was found in a Korean sample. Loss-of-function mutations in PMS2 are associated with mismatch repair defects that lead to a higher risk of cancer development. In a separate study of gall bladder cancer, we found the same essential splice site PMS2 mutation (chromosome 7:6043690C>G) in a Korean patient whose gall bladder cancer exhibits microsatellite instability (E.W.S. and S. Seshagiri, manuscript in preparation). Identification of genetic variants that affect drug efficacy and safety through the alteration of pharmacokinetics enables application of individualized treatment.

Variation in drug responses are generally recognized and recommendations for dosing are sometimes guided by apparent or self-reported population identity despite the lack of a rigorous pharmacogenomic basis. We assessed the allele frequencies of key pharmacogenomic variants in our dataset to identify inter-population differences that have potential implications on drug testing and treatment (Fig. 3e, Supplementary Table 4g and Supplementary Information 13).

Carbamazepine, clopidigrel, peginterferon and warfarin showed the largest variation between populations in predicted adverse drug responses with groups ranging from 0 and 100 predicted adverse drug
variants (recessive (AR) or other (unknown) classification as per OMIM. A number of those designated disease-causing in the Human Gene Mutation Database (HGMD) or pathogenic in ClinVar. Autosomal-dominant (AD) or autosomal-recessive (AR) or other (unknown) classification per OMIM. A number of variants (n = 152). The centre line indicates the median; box limits show the upper and lower quartiles; whiskers extend 1.5× the interquartile range.

In the GAsP study dataset at high frequency and were reclassified (Supplementary Table 4d). c. Frequency of β-thalassaemia variant (chromosome 11:5248155 c.92+5G>C) across Asia shows a geographical enrichment. MAF in South Asia is 1.4%. NA, not available. d. Novel cancer-predisposing variants identified in the GenomeAsia dataset. e. Population-specific probabilities of adverse drug reactions predicted from the aggregate allele frequencies of known variants associated with response to the indicated drugs.

Our analysis revealed IBD scores of 1.465 and 0.817 for Finnish and British groups, consistent with previous analyses45. The IBD score of all of the groups was normalized relative to the Finnish group (Fig. 4a and Supplementary Information 12). Our study includes many groups with small population sizes and it is expected that endogamy paired with small population size will greatly increase IBD scores. We found that indigenous and tribal groups had IBD scores that were skewed upwards from non-tribal groups (Fig. 4b). Notably, we found that a number of Asian groups with large urban populations have IBD scores above or close to that of the Finnish population. For example, samples from an outpatient hospital in Chennai, a city with a census size of 9 million, had an IBD score that was approximately 1.3 times greater than the score for the Finnish group.

**Founder populations**

Population bottlenecks produce strong founder effects and increased rates of recessive disease. In populations with strong founder effects, the loss-of-function variant frequency spectrum is skewed higher, greatly increasing power of association and providing unique advantages for the identification of genes associated with both rare and complex diseases43,44. We followed the approach described in a previous study on south Asian populations to characterize the degree to which genomic segments are inherited as identical by descent (IBD) in population groups in our dataset45.

**Human knockouts**

Homozygous loss-of-function alleles found in humans give us the opportunity to assess the phenotypic effect of specific gene loss and can provide important information about opportunities for treating disease46,47. To assess the contents of our dataset, we examined high-confidence protein-truncating variants (PTVs). We found 17,566 PTVs with at least 1 PTV in approximately 43% of all protein-coding genes (n = 8,766; Fig. 4c). Among the PTVs, most were heterozygous variants.

---

Fig. 3 | Disease-relevant variant discovery. a. Filtering using the GAsP dataset improves candidate variant discovery by removing population specific variants (n = 152). The centre line indicates the median; box limits show the upper and lower quartiles; whiskers extend 1.5× the interquartile range. b. Allele count (AC) and frequency distribution of variants in the GAsP dataset that are designated disease-causing in the Human Gene Mutation Database (HGMD) or pathogenic in ClinVar. Autosomal-dominant (AD) or autosomal-recessive (AR) or other (unknown) classification per OMIM. A number of variants (n = 37) that had previously been reported to be pathogenic are found.

---

responses. For example, the HLA-B*15:02 variant, associated with risk for development of Steven Johnson syndrome48 in patients treated with carbamazepine was found to occur at an increased frequency in Austronesian language-speaking populations from southeast Asia (for example, 63% in the Mentawai of West Sumatra; 46.6% in the Nias of North Sumatra) compared with other groups (Supplementary Information 13). There are roughly 400 million individuals who belong to Austronesian groups that are at increased risk for carbamazepine sensitivity, including the vast majority of the people from Indonesia, Malaysia and the Philippines.

Fig. 3 | Disease-relevant variant discovery. a. Filtering using the GAsP dataset improves candidate variant discovery by removing population specific variants (n = 152). The centre line indicates the median; box limits show the upper and lower quartiles; whiskers extend 1.5× the interquartile range. b. Allele count (AC) and frequency distribution of variants in the GAsP dataset that are designated disease-causing in the Human Gene Mutation Database (HGMD) or pathogenic in ClinVar. Autosomal-dominant (AD) or autosomal-recessive (AR) or other (unknown) classification per OMIM. A number of variants (n = 37) that had previously been reported to be pathogenic are found.
Fig. 4 | Founder effects and homozygous loss of function. a, IBD scores across different population groups are shown for 96 ethnicities (1,417 samples) across global regions. The scores given in the figure are relative ratios compared to that of the Finnish group. b, Violin plot showing IBD scores in 29 tribal groups and 25 non-tribal groups consisting of 293 and 336 samples, respectively. The centre line indicates the median; box limits show 1.5× the interquartile range.

c. Proportion of genes with at least one high-confidence PTV. d, Proportion of novel, known, heterozygous and homozygous PTVs in the GasP dataset. e, Pie chart of novel homozygous PTVs plotted by region (inner circle) and population group (outer circle). Groups with less than two PTVs were grouped as other. f, Novel homozygous PTV Q2010* (green) found in ABCA7 localizes to the C-terminal ABC domain. Previously reported PTVs are shown in grey.
Article

GenomeAsia100K Consortium

Jeffrey D. Wall1,4, Eric W. Stawiski2,4,47, Aakrosh Ratan4,47, Hie Lim Kim4,47, Changhoon Kim4,47, Ravi Gupta4,47, Kushal Suryamohan5, Elena S. Gusareva6, Rikky Wenang Purbojati6, Tushar Bhangale7, Vladimír Stepánov7,8,9, Vladimir Kharkov7,8,9, Markus S. Schröder10, Vedam Ramprasad11, Jennifer Tom12, Steffen Durinck13, Xixin Bei14, Jiani Li14, Joseph Guillory14, Sameer Phake15, Anubhav Basu16, Jeremy Stinson17, Sandhya Nair18, Sivasankar Malaiycham18, Nidhan K. Biswas19, John C. Chambers19, Keith C. Cheng19, Joyner T. George19, Seik Soon Khor19, Jong Il Kim19,20, Belong Cho20, Ramesh Menon20, Thirumalasetti Sattibabu20, Akshi Bassi20, Manjari Deshmukh21, Anjali Verma22, Vivek Gopalan22, Jong-Heon Shin23,24, Mahesh Pratapneni25, Sam Santhosh26, Katsushi Tokunaga27,28,29, Badrul M. Md-Zain29,30, Kok Gan Chan26,30, Madasamy Parani27, Purushothaman Natarajan27, Michael Hauser28,29, R. Rand Allingham30,31,32,33,34,35, Cecilia Santiago-Turra36, Arkasubhra Ghosh30,31,32,33,34,35, Santosh Gopi Krishna Gadoth36, Christian Fuchsberger37,38,39, Lukas Forer39, Sebastian Schoenhammer40,41,42,43,44, Herawati Sudoyo40,41, J. Stephen Lansing5, Jonathan Friedlaender36, George Koh32, Murray P. Cox32, Michael Hammer32, Tatiana Karafet45, Khai C. Anh46,47, Syed Q. Mehdi46,47, Venkatesan Radha46,47, Viswanathan Mohan46,47, Partha P. Majumder46,47, Somasekar Seshagiri46,47, Jeong-Sun Seo48,49,50, Stephan C. Schuster48,49,50 & Andrew S. Peterson48,49,50

1Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. 2Department of Molecular Biology, Genentech, South San Francisco, CA, USA. 3Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA, USA. 4Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. 5The Asian School of the Environment, Nanyang Technological University, Singapore, Singapore. 6Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, Singapore. 7Bioinformatics Institute, Macrogen, Seoul, South Korea. 8Precision Medicine Center, Seoul National University Bundang Hospital, Gyeonggi-do, South Korea. 9MedGenome Labs, Bengaluru, India. 10Department of Human Genetics, Genentech, South San Francisco, CA, USA. 11Institute of Medical Genetics, Tomsk National Medical Research Center, Tomsk, Russian Federation. 12Russian Academy of Sciences, Tomsk, Russian Federation. 13Tomsk State University, Tomsk, Russian Federation. 14National Institute of BioMedical Genomics, Netaji Subhas Sanatorium, Kalyani, India. 15Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. 16Division of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea. 17Department of Family Medicine, Seoul National University Hospital, Seoul, South Korea. 18Department of Family Medicine, Seoul National University Hospital, Seoul, South Korea. 19Precision Medicine Institute, Macrogen, Gyeonggi-do, South Korea. 20GenomeAsia100K Consortium, Genomika-Korea, Seoul, South Korea. 21Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea. 22Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul, South Korea. 23Department of Family Medicine, Seoul National University Hospital, Seoul, South Korea. 24Dr. Mohan’s Diabetes Specialities Centre, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India. 25Department of Molecular Biology, Genentech, South San Francisco, CA, USA. 26Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, Bangi, Malaysia. 27WHO Collaborating Centre for Medical Genetics, University of Tokyo, Tokyo, Japan. 28Department of Environment and Natural Resource Science, School of Environment and Natural Resource Science, Faculty of Science and Technology, University Kebangsaan Malaysia, Bangi, Malaysia. 29Department of Medical Genetics, Institute of Genetic Medicine, National Institute of BioMedical Research, Bengaluru, India. 30Department of Ophthalmology, Duke University Medical Center, Durham, NC, USA. 31Department of Pathology and Jake Vedam Ramprasad, Jennifer Tom, Steffen Durinck, Xixin Bei, Jiani Li, Joseph Guillory, Sameer Phake, Anubhav Basu, Jeremy Stinson, Sandhya Nair, Sivasankar Malaiycham, Nidhan K. Biswas, John C. Chambers, Keith C. Cheng, Joyner T. George, Seik Soon Khor, Jong Il Kim, Belong Cho, Ramesh Menon, Thirumalasetti Sattibabu, Akshi Bassi, Manjari Deshmukh, Anjali Verma, Vivek Gopalan, Jong-Heon Shin, Mahesh Pratapneni, Sam Santhosh, Katsushi Tokunaga, Badrul M. Md-Zain, Kok Gan Chan, Madasamy Parani, Purushothaman Natarajan, Michael Hauser, R. Rand Allingham, Cecilia Santiago-Turra, Arkasubhra Ghosh, Santosh Gopi Krishna Gadoth, Christian Fuchsberger, Lukas Forer, Sebastian Schoenhammer, Herawati Sudoyo, J. Stephen Lansing, Jonathan Friedlaender, George Koh, Murray P. Cox, Michael Hammer, Tatiana Karafet, Khai C. Anh, Syed Q. Mehdi, Venkatesan Radha, Viswanathan Mohan, Partha P. Majumder, Somasekar Seshagiri, Jeong-Sun Seo, Stephan C. Schuster, Andrew S. Peterson. *e-mail: ppm1@nibmg.ac.in; sekar@sgrf.org; jeongsun@snu.ac.kr; stephan.c.schuster@gmail.com; peterson.andrew@genomeasia100k.org
Methods

Data reporting
No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to the allocation during analysis.

Samples
We accessed publicly available high-coverage, whole-genome FASTQ files from previous studies of human genetic variation and combined these with 1,267 high-coverage genomes generated as part of this project. Full details on the samples chosen for sequencing and the informed consent processes for these samples can be found in Supplementary Information 1. We restricted our analyses to genomes generated using Illumina short-read sequencing technology.

Whole-genome sequencing
Whole-genome sequencing libraries were prepared using standard protocols (Illumina) and sequenced on Illumina HiSeq 2500/4000 or X10 machines. We obtained paired-end (2 × 100 bp or 2 × 150 bp) for each sample.

Filtering, alignment and variant calling
We aligned the Illumina short-read sequences to the GRCh37+decov reference genome with BWA-mem using the default parameters. Putative PCR duplicates were flagged using SAMBLASTER. The SAM outputs were converted to BAM format, and sorted by chromosomal coordinates using Sambamba, and all BAM files for the same samples were merged.

The sex of the samples was inferred from the coverage of the autosomes and the sex chromosomes, and confirmed from the submitted metadata with the samples. All samples that had an average coverage less than 20-fold or for which we found a difference in the inferred and reported sex were removed from further analysis. We used verifyBamID to identify contamination using the chip-free mode and samples for which swaps or contamination was identified were removed from subsequent analyses. A contamination level of 3% was used as a cut-off, and this left us with 1,739 samples that were used for all downstream analyses.

We identified the single-nucleotide substitutions and small indels variants in the 1,739 samples using the reference model (gVCF-based) workflow for joint analysis in GATK. Variants were called individually in each sample using the HaplotypeCaller in ‘--ERC GVCF’ mode to produce a record of genotype likelihoods and annotations at each site in the genome. Multi-allelic variants are reported in the GenomeAsia browser but were not included in the analysis. A gVCF file was created for every sample, and a subsequent joint genotyping analysis of all gVCFs was done to identify the variants in the cohort. We followed the GATK-recommended best practices for variant recalibration to create a final VCF file and recalibrated the variants to select 99% of the true sites from the training set for VQS.R. The VCF files were zipped using bgzip and indexed using tabix.

Identification of first-degree relative pairs
Several of the reported analyses require filtering to remove related samples. We used KING to identify such first-degree relative pairs. We first used vcftools and plink to convert the VCF file into the required input format for KING. The estimated kinship coefficient was restricted to 0.177–0.354 as described in the KING manual to identify the first-degree relative pairs, and the results were confirmed from the submitted metadata. The number of unrelated samples by country-of-origin is shown in Supplementary Table 1.1.

Quantifying population structure and changes in population size
We restricted our attention to 7,966,132 autosomal markers (that is, SNPs) with MAF ≥ 0.01 and call rate ≥ 98%. In some analysis, severe linkage disequilibrium pruning was applied as follows: sliding windows of size 50 (that is, the number of markers used for linkage disequilibrium testing at a time) and window increments of 5 markers; for any pair of SNPs in an window, the first marker of the pair was discarded if r > 0.2. After linkage disequilibrium pruning, 1,089,227 SNPs were retained for analysis. All data-filtrering procedures were conducted in PLINK v.1.9.

Analyses of population structure was performed using the quality-control-positive linkage-disequilibrium-pruned set of 1,089,227 autosomal SNPs. Principal component analysis (PCA) was conducted across all available populations in EIGENSTRAT v.6.1.4. Results were visualized in Tableau v.9.3. We applied unsupervised hierarchical clustering of individuals using the maximum likelihood method implemented in ADMIXTURE v.1.3.0 using default input parameters. The ‘--cv’ flag was adopted to perform the cross-validation procedure and to calculate the optimal k value.

We used MSMC to estimate changes in population size and split times. This analysis used two different phased genomes datasets (using Shapeit v.2 and Eagle 2). The details for the phasing are described in Supplementary Information 4. Chromosome 6 was excluded from the analysis owing to possible phasing errors in the HLA region. We used four haplotypes (two individual genomes) for estimating changes in population size in a population and eight haplotypes (two genomes from each of a pair of populations) for the estimation of population split times. We assumed a mutation rate of μ = 1.25 × 10−8 per site per generation and an average generation time of 29 years, as in previous studies.

Comparison with 1000 Genomes Project genotype calls
We filtered the variant calls to include only biallelic SNPs with <10% missing genotype calls, that were within the 1000 Genomes Project mask (available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.strict_mask.whole_genome.bed). Then, for each of the 119 overlapping samples considered individually, we calculated variant discordance rates for those filtered SNPs that (1) had a genotype call in both the 1000 Genomes Project data and the GAsP data; and (2) had a ‘variant’ call (that is, a non-homozygous reference genotype call) in at least one of the datasets. These discordance rates were then stratified by the estimated MAF in the GAsP dataset.

Patterns of allele sharing
We used a parsimony-based analysis of allele sharing that focused on SNPs that were not present in sub-Saharan Africans or in archaic humans (further details are provided in Supplementary Information 8).

Archaic admixture
We used a method similar to the ‘enhanced’ D-statistic approach to estimate levels of Neanderthal and Denisovan ancestry in each non-African sample. The estimates were calibrated assuming 0% Denisovan ancestry in the British population, 4% Denisovan ancestry in the Papuan population and 2% Neanderthal ancestry in the British population (full details are provided in Supplementary Information 9).

Determination of high-quality variants for medically related analyses
High-quality variants were defined as variants that (1) had a read-depth ≥ 5 and genotype-quality ≥ 20; (2) were contained in the high-confidence regions as described by Genome in a Bottle (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2_GRCh37/SupplementalFiles/HG001_GRCh37_GIAB_highconf_CG-llFB-IllGATKH-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf.bed) and (3) passed the gnomAD Filter. Variant annotation was carried out using Snpeff (v.4.1).

IBD scores
Groups with at least two samples were considered for analysis. We restricted our analysis to genomic regions with high-confidence calls...
and removed related samples based on reported relationship, kinship, PCA and IBD analyses. The scores given in the figure are relative ratios compared to that of the Finnish group.

PTVs
PTVs are defined as high-quality variants that were annotated as having a strong impact on the protein (such as frameshifts, essential splice sites or premature stop codons). We restricted calls to high-confidence regions determined by Genome in a Bottle as described above and filtered for high-confidence PTVs using the LOFTEE program\(^6^9\). We used a similar strategy for additional filtering of variants as proposed previously\(^6^8\) and flagged variants with ≤7 reads covering the variant site: ≤80% of reads had the variant, were not in the bottom 1 percentile of phylOp or gerpRS\(^6^8\) scores and for which the affected transcripts made up less than 50% of all expression as specified by GTEx.

Enriched medically relevant variants
We compared variant allele counts for Asian and Oceania samples from the GenomeAsia cohort to allele counts present in non-Asian gnomAD samples (European (non-Finnish), European (Finnish), Latino, African or other) for variants found in a set of 124 medically relevant genes. The genes used were 115 genes used for prenatal screening\(^6^6\) as well as the cancer-associated genes BRCA1, BRCA2, TP53, MENI, MLH1, MSH2, MSH6, PMS1 and PMS2A. A Fisher's exact test was used to calculate variations that were significantly overrepresented in the GenomeAsia subsamples and corrected for multiple testing using the Bonferroni method. We further accessed variants for these genes that had not previously been reported. All variants were further filtered as being damaging as determined by having a high impact on the protein (stop codon, essential splice site or frameshift mutation) or were predicted to be damaging by the Polyphen2 program. A cumulative comparison of allele counts for all over-represented and novel variants was performed and compared to non-Asian gnomAD to calculate a P value, odds ratio and relative difference in cumulative allele frequency (GenomeAsia cumulative allele frequency minus gnomAD non-Asian allele frequency). Reported P values were corrected for multiple testing using the Bonferroni method.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
For each variant, summary data for genotype quality, allele depth and population-specific allele counts were calculated before removing all genotype data. This dataset is available without requirement for login or other form of restriction for browsing or for download at https://browser.genomeasia100k.org. Individual level VCF data files representing the 1,180 newly sequenced genomes from 74 populations are freely available to any qualified investigator without restriction. Chinese samples sequenced were from Corriell cell lines (https://coriell.org/). The goal of the GenomeAsia100K consortium is to facilitate and accelerate genetic studies in Asian populations by coordinating sequencing efforts among its members. To achieve this goal, we are committed to continuing to make data publicly available and accessible. As data are contributed to the consortium by individual members, it will be made immediately available in summary form or as imputation reference panels where appropriate. Data will be made available in individual form wherever possible and not limited by the bounds of informed consent, national privacy laws and regulations, or other external restrictions that may apply.

Acknowledgements
We thank the many individuals from all across Asia who gave blood samples for scientific research and the many individuals who supported the sample collection. The computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

Author contributions
J.D.W., E.W.S., A.R. Aras, K.C.C., M. Pratapneni, S. Santhosh, H.S., J.S.L., P.P.M., J.-S.S., S.C.S., S. Seshagiri and A.S.P. designed the study. S. Seshagiri, A.S.P., S.C.S., V. Ramprasad, J.G.S. and J.T.G. produced the sequencing data. S. Seshagiri, S.P., J.D.W., E.W.S., R.W.P. and A.R. carried out the data processing and quality control. P.P.M., K.C.C., J.-S.S., S.C.S., J.T.G., J.S.L., S.-S.N., S.M., J.T.G., S.K., S.G.K.G., K.G.C., J.-I.K., C.K., B.C., B.M., B.M.M.-Z., K.T., M. Pratapneni, V. Radha and S. Seshagiri, M. Hauser, R.R.A., A.G., M.P.C., J.F., M. Hammer, T.K., K.C.A., S.Q.M., V. M., V. Radha and S. Seshagiri coordinated, collected and/or provided samples. C.F., L.F. and S. Schoener generated the imputation server: P.P.M., S. Seshagiri, J.D.W., E.W.S., A.R., A.S.P., H.L.K., R.G., K.S., E.S.G., T.B., V.K., V.S., M.S.S., J.T., S.D., Q.B., J.L., N.K.B., R.M., T.S., A.V., V.G., A. Bassi, A. Basu, C.K. and M.D. carried out analyses. J.D.W., S. Seshagiri, E.W.S. and A.S.P. wrote the paper.

Competing interests
A.S.P., E.W.S., S. Seshagiri, T.B., J.G.T., J.T., J. Stinson, Q.B., M.S.S. and K.S. were employees of Genentech at the time this work was carried out. S. Santhosh, A.V., M. Pratapneni, V. Ramprasad, S.P., R.G., S.N., S.M., T.S., V.G., J.T.G., M.D. and S.P. are employees and/or have equity in MedGenome. C.K., J.-S.S. and J.Y.S. are employees of Macrogen.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1793-z. Correspondence and requests for materials should be addressed to S. Seshagiri, J.-S.S., S. Schoener and A.S.P.

Peer review information
Nature thanks Rasmus Nielsen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Diversity and divergence times of GAsP samples.

(a) PCA plot of study samples. Africa (AFR), n = 102; West Eurasia (WER), n = 111; South Asia (SAS), n = 642; Southeast Asia (SEA), n = 162; Oceania (OCE), n = 68; Northeast Asia (NEA), n = 346; Americas (AMR), n = 26. The samples included in each of these geographically defined groups are described in Supplementary Table 1a.

(b) MSMC cross-coalescence rates showing divergence time estimates between different groups. The point estimate of the date was given at which 25%, 50% and 75% of lineages in the pair of populations have coalesced into a common ancestral population.
Extended Data Fig. 2 | Characteristics of GAsP SNPs and indels.
a, b, Comparison of all GAsP variants (a) or coding variants (b) with gnomAD, ExAC, 1000 Genomes, ESP and dbSNP data as a function of the MAF within the GAsP dataset. c, d, The number and lengths of small indels in the genome (c) or coding regions (d). e–h, Proportion of non-coding (e, g) or coding (f, h) indels that were singletons (e, f) or rare (allele frequency of $<0.1\%$; g, h).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection no software was used

Data analysis

- BWA version 0.7.13 (https://github.com/lh3/bwa);
- SAMBLASTER version 0.1.22 (https://github.com/GregoryFaust/samblaster)
- Sambamba version 0.6.1 (https://github.com/lomereiter/sambamba)
- BAMreport version 0.0.2; (https://github.com/aakrosh/BAMreport)
- verifyBamID version 1.1.3 (http://genome.sph.umich.edu/wiki/VerifyBamID);
- GATK version 3.5 (https://software.broadinstitute.org/gatk/);
- vcfanno version 0.1.0-dev (https://github.com/brentp/vcfanno);
- htslib version 1.3.1-64-g74bcfd7 (https://github.com/samtools/htslib);
- vcftools version 0.1.14 (https://vcftools.github.io/index.html);
- plink version 1.90b3.40 (http://zzz.bwh.harvard.edu/plink/);
- king version 1.4 (http://people.virginia.edu/~wc9c/KING/);
- rtg-tools version 3.7 (https://github.com/RealTimeGenomics/rtg-tools);
- Shapeit v2 (Delaneau et al, 2012);
- ex-tractPIRs (Delaneau et al, 2013);
- Eagle2 algorithm (Loh et al. 2016), version 2.3;
- generate_multihetsep.py, downloaded from https://github.com/stschiff/msmc-tools;
- Admixture v.1.3.0 (Alexander et al, 2009);
- EIGENSTRAT v.6.1.4 (Price et al, 2006);
- Selscan v. 1.1.0 (Szpiech and Hernandez 2014);
- BEAST v.1.8.4 (Drummond et al. 2012);
- PLINK v1.9

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

For each variant, summary data for genotype quality, allele depth and population specific allele counts were calculated before removing all genotype data. This data set is available without requirement for login or other form of restriction for browsing or for download at (https://browser.genomeasia100k.org). Individual level VCF data files representing 1,180 newly sequenced genomes from individuals in 74 population groups are freely available to any qualified investigator without restriction. Chinese samples sequenced were from Coriell cell lines and are not subject to the Chinese regulation. The data are available from the European Genome Archive (EGA) under accession number EGAS00001002921.

The procedure for accessing individual level data is as follows:
Access forms obtained from the GenomeAsia website (https://browser.genomeasia100k.org), once filled out and returned to dataaccess@genomeasia100k.org will undergo administrative review and instructions for download will be returned to the requestor. Access to individual level data from Malaysian samples are subject to additional restrictions.
The complete data set of sequences of unrelated individuals (1,667 samples) has been phased and can be used for imputation through the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html).
The goal of the GenomeAsia100k consortium is to facilitate and accelerate genetic studies in Asian populations by coordinating sequencing efforts amongst its members. To achieve this goal we are committed to continuing to make data publicly available and accessible. As data is contributed to the consortium by individual members it will be made immediately available in summary form or as imputation reference panels where appropriate. Data will be made available in individual form wherever possible and not limited by the bounds of informed consent, national privacy laws and regulations, or other external restrictions that may apply.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [] Behavioural & social sciences
- [] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No statistical methods were used to predetermine sample size. and the investigators were not blinded to the allocation during analysis. |
| Data exclusions | data was not excluded unless it failed essential QC metrics |
| Replication | results were not externally replicated |
| Randomization | The experiments were not randomized. |
| Blinding | Investigators were not blinded to the allocation during analysis. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
| --- | --- |
| [x] | Antibodies |
| [x] | Eukaryotic cell lines |
| [x] | Palaeontology |
| [x] | Animals and other organisms |
| [x] | Human research participants |
| [x] | Clinical data |

Methods

| n/a | Involved in the study |
| --- | --- |
| [x] | ChIP-seq |
| [x] | Flow cytometry |
| [x] | MRI-based neuroimaging |
## Human research participants

Policy information about [studies involving human research participants](https://www.nature.com/articles/nrg.2018.112#nrg.2018.112).

| Population characteristics | 200 populations groups were included in our study and study participants included equal numbers of both genders |
|-----------------------------|----------------------------------------------------------------------------------------------------------|
| Recruitment                 | Participants were recruited based on self and external identification as member of a specific population groups |
| Ethics oversight            | Nanyang Technological University institutional review board (IRB- 2014-12-011)                           |

Note that full information on the approval of the study protocol must also be provided in the manuscript.