Title: The landscape of tolerated genetic variation in humans and primates

Authors: Hong Gao[1]†, Tobias Hamp[1]†, Jeffrey Ede[1], Joshua G. Schraiber[1], Jeremy McRae[1], Moriel Singer-Berk[2], Yanshen Yang[1], Anastasia Dietrich[1], Petko Fiziev[1], Lukas Kuderna[1,3], Laksshman Sundaram[1], Yibing Wu[1], Aashish Adhikari[1], Yair Field[1], Chen Chen[1], Serafim Batzoglou[1]‡, Francois Aguet[1], Gabrielle Lemire[2,4], Rebecca Reimers[4], Daniel Balick[5], Mareike C. Janiak[6], Martin Kuhlwilm[3,7,8], Joseph D. Orkin[3,9], Shivakumara Manu[10,11], Alejandro Valenzuela[3], Juraj Bergman[12,13], Marjolaine Rouselle[12], Felipe Ennes Silva[14,15], Lidia Agueda[16], Julie Blanc[16], Marta Gut[16], Dorien de Vries[6], Ian Goodhead[6], R. Alan Harris[17], Muthuswamy Raveendran[17], Axel Jensen[18], Idriss S. Chuma[19], Julie Horvath[20,21,22,23,24], Christina Hvilsom[25], David Juan[3], Peter Frandsen[25], Fabiano R. de Melo[26], Fabricio Bertuol[27], Hazel Byrne[28], Iracilda Sampaio[29], Izeni Farias[27], João Valsecchi do Amaral[30,31,32], Mariluce Messias[33,34], Maria N. F. da Silva[35], Mihir Trivedi[11], Rogerio Rossi[36], Tomas Hrbek[27,37], Nicole Andriaholinirina[38], Clément J. Rabarivola[38], Alphonse Zaramody[38], Clifford J. Jolly[39], Jane Phillips-Conroy[40], Gregory Wilkerson[41]§, Christian Abee[42], Joe H. Simmons[41], Eduardo Fernandez-Duque[42,43], Sree Kanthaswamy[44], Fekadu Shiferaw[45], Dongdong Wu[46], Long Zhou[47], Yong Shao[46], Guojie Zhang[47,48,49,50,51], Julius D. Keyyu[52], Sascha Knauf[53], Minh D. Le[54], Esther Lizano[3,55], Stefan Merker[56], Arcadi Navarro[3,57,58,59], Thomas Batallion[12], Tilo Nadler[60], Chiea Chuen Khor[61], Jessica Lee[62], Patrick Tan[61,63,64], Weng Khong Lim[63,64,65], Andrew C. Kitchener[66,67], Dietmar Zinner[68,69,70], Ivo Gut[16,71], Amanda Melin[72,73], Katerina Guschanski[18,74], Mikkel Heide Schierup[12], Robin M. D. Beck[6], Govindhaswamy Umapathy[10,11], Christian Roos[75], Jean P. Boubli[6], Monkol Lek[76], Shamil Sunyaev[77,5], Anne O'Donnell[2,4,78], Heidi Rehm[2,79], Jinbo Xu[1,80], Jeffrey Rogers[17]*¶, Tomas Marques-Bonet[3,16,55,57]*, Kyle Kai-How Farh[1]*

Affiliations:

[1]Illumina Artificial Intelligence Laboratory, Illumina Inc.; Foster City, California, 94404, USA.

[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard; Boston, Massachusetts, 02142, USA.

[3]Institute of Evolutionary Biology (UPF-CSIC); PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain.

[4]Division of Genetics and Genomics, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School; Boston, Massachusetts, 02115, USA.

[5]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School; Boston, Massachusetts, 02115, USA.

[6]School of Science, Engineering & Environment, University of Salford; Salford, M5 4WT, United Kingdom.

[7]Department of Evolutionary Anthropology, University of Vienna; Djerassiplatz 1, 1030, Vienna, Austria.

[8]Human Evolution and Archaeological Sciences (HEAS), University of Vienna; 1030, Vienna, Austria.

[9]Département d'anthropologie, Université de Montréal; 3150 Jean-Brillant, Montréal, QC, H3T 1N8, Canada.

[10]Academy of Scientific and Innovative Research (AcSIR); Ghaziabad, 201002, India.

[11]Laboratory for the Conservation of Endangered Species, CSIR-Centre for Cellular and Molecular Biology; Hyderabad, 500007, India.
[12]Bioinformatics Research Centre, Aarhus University; Aarhus, 8000, Denmark.
[13]Section for Ecoinformatics & Biodiversity, Department of Biology, Aarhus University; Aarhus, 8000, Denmark.
[14]Research Group on Primate Biology and Conservation, Mamirauá Institute for Sustainable Development; Estrada da Bexiga 2584, Tefé, Amazonas, CEP 69553-225, Brazil.
[15]Faculty of Sciences, Department of Organismal Biology, Unit of Evolutionary Biology and Ecology, Université Libre de Bruxelles (ULB); Avenue Franklin D. Roosevelt 50, 1050, Brussels, Belgium.
[16]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST); Baldiri i Reixac 4, 08028, Barcelona, Spain.
[17]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine; Houston, Texas, 77030, USA.
[18]Department of Ecology and Genetics, Animal Ecology, Uppsala University; SE-75236, Uppsala, Sweden.
[19]Tanzania National Parks; Arusha, Tanzania.
[20]North Carolina Museum of Natural Sciences; Raleigh, North Carolina, 27601, USA.
[21]Department of Biological and Biomedical Sciences, North Carolina Central University; Durham, North Carolina , 27707, USA.
[22]Department of Biological Sciences, North Carolina State University; Raleigh, North Carolina , 27695, USA.
[23]Department of Evolutionary Anthropology, Duke University; Durham, North Carolina , 27708, USA.
[24]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
[25]Copenhagen Zoo; 2000 Frederiksberg, Denmark.
[26]Universidade Federal de Viçosa; Viçosa, 36570-900, Brazil.
[27]Universidade Federal do Amazonas, Departamento de Genética, Laboratório de Evolução e Genética Animal (LEGAL); Manaus, Amazonas, 69080-900, Brazil.
[28]Department of Anthropology, University of Utah; Salt Lake City, Utah, 84102, USA.
[29]Universidade Federal do Para; Guamá, Belém - PA, 66075-110, Brazil.
[30]Research Group on Terrestrial Vertebrate Ecology, Mamirauá Institute for Sustainable Development; Tefé, Amazonas, 69553-225, Brazil.
[31]Rede de Pesquisa para Estudos sobre Diversidade, Conservação e Uso da Fauna na Amazônia – RedeFauna; Manaus, Amazonas, 69080-900, Brazil.
[32]Comunidad de Manejo de Fauna Silvestre en la Amazonía y en Latinoamérica – ComFauna; Iquitos, Loreto, 16001, Peru.
[33]Universidade Federal de Rondonia; Porto Velho, Rondônia, 78900-000, Brazil.
[34]PPGREN - Programa de Pós-Graduação "Conservação e Uso dos Recursos Naturais and BIONORTE - Programa de Pós-Graduação em Biodiversidade e Biotecnologia da Rede BIONORTE, Universidade Federal de Rondonia; Porto Velho, Rondônia, 78900-000, Brazil.
[35]Instituto Nacional de Pesquisas da Amazonia; Petrópolis, Manaus - AM, 69067-375, Brazil.
[36]Universidade Federal do Mato Grosso; Boa Esperança, Cuiabá - MT, 78060-900, Brazil.

[37]Department of Biology, Trinity University; San Antonio, Texas, 78212, USA.
[38]Life Sciences and Environment, Technology and Environment of Mahajanga, University of Mahajanga; Mahajanga, 401, Madagascar.
[39]New York University; New York City, 10012, USA.
[40]Washington University in St. Louis; St. Louis, Missouri,63130, USA.
[41]Keeling Center for Comparative Medicine and Research, MD Anderson Cancer Center; Houston, Texas, 77030, USA.
[42]Yale University; New Haven, Connecticut, 06520, USA.
[43]Universidad Nacional de Formosa, Argentina Fundacion ECO, Formosa, Argentina.
[44]Arizona State University; Tempe, Arizona, 85281, USA.
[45]Hawassa University; Hawassa, 005, Ethiopia.
[46]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences; Kunming, Yunnan, 650223, China.
[47]Center for Evolutionary & Organismal Biology, Zhejiang University School of Medicine, Hangzhou, 310058, China.
[48]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen; Copenhagen, DK-2100, Denmark.
[49]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China
[50]Liangzhu Laboratory, Zhejiang University Medical Center; 1369 West Wenyi Road, Hangzhou, 311121, China
[51]Women's Hospital, School of Medicine, Zhejiang University; 1 Xueshi Road, Shangcheng District, Hangzhou, 310006, China
[52]Tanzania Wildlife Research Institute (TAWIRI), Head Office; P.O.Box 661, Arusha, Tanzania.
[53]Institute of International Animal Health/One Health, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health; 17493 Greifswald - Isle of Riems, Germany.
[54]Department of Environmental Ecology, Faculty of Environmental Sciences, University of Science and Central Institute for Natural Resources and Environmental Studies, Vietnam National University; Hanoi, 100000, Vietnam.
[55]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain; Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain
[56]Department of Zoology, State Museum of Natural History Stuttgart; 70191 Stuttgart, Germany.
[57]Institució Catalana de Recerca i Estudis Avançats (ICREA) and Universitat Pompeu Fabra, Pg. Luís Companys 23, Barcelona, 08010, Spain.
[58]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology; Av. Doctor Aiguader, N88, Barcelona, 08003, Spain.
[59]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation; C. Wellington 30, Barcelona, 08005, Spain.
[60]Cuc Phuong Commune; Nho Quan District, Ninh Binh Province, 430000, Vietnam.
[61]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore.
[62]Mandai Nature; 80 Mandai Lake Road, Singapore 729826, Republic of Singapore.
[63]SingHealth Duke-NUS Institute of Precision Medicine (PRISM); Singapore 168582, Republic of Singapore.

[64]Cancer and Stem Cell Biology Program, Duke-NUS Medical School; Singapore 168582, Republic of Singapore.
[65]SingHealth Duke-NUS Genomic Medicine Centre; Singapore 168582, Republic of Singapore.
[66]Department of Natural Sciences, National Museums Scotland; Chambers Street, Edinburgh, EH1 1JF, UK.
[67]School of Geosciences, University of Edinburgh; Drummond Street, Edinburgh, EH8 9XP, UK.
[68]Cognitive Ethology Laboratory, Germany Primate Center, Leibniz Institute for Primate Research; 37077 Göttingen, Germany.
[70]Department of Primate Cognition, Georg-August-Universität Göttingen; 37077 Göttingen, Germany.
[71]Universitat Pompeu Fabra, Pg. Luís Companys 23, Barcelona, 08010, Spain.
[72]Leibniz Science Campus Primate Cognition; 37077 Göttingen, Germany.
[73]Department of Anthropology & Archaeology and Department of Medical Genetics
[74]Alberta Children's Hospital Research Institute; University of Calgary; 2500 University Dr NW T2N 1N4, Calgary, Alberta, Canada.
[75]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh; Edinburgh, EH8 9XP, UK.
[76]Gene Bank of Primates and Primate Genetics Laboratory, German Primate Center, Leibniz Institute for Primate Research; Kellnerweg 4, 37077 Göttingen, Germany.

[77]Department of Genetics, Yale School of Medicine; New Haven, Connecticut, 06520, USA.

[78]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, 02115, USA.

[79]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School; Boston, Massachusetts, 02115, USA.

[80]Toyota Technological Institute at Chicago; Chicago, Illinois, 60637, USA.

† These authors contributed equally to this work

*Corresponding authors. Email: tomas.marques@upf.edu, jr13@bcm.edu, kfarh@illumina.com

‡ Current affiliation: Seer, Inc., Redwood City, California, 94065, USA

§ Current affiliation: Department of Clinical Science, College of Veterinary Medicine, North Carolina State University, Raleigh, North Carolina, 27606, USA

¶ Current affiliation: Wisconsin National Primate Research Center, Madison, Wisconsin, 53715, USA

Abstract: Personalized genome sequencing has revealed millions of genetic differences between individuals, but our understanding of their clinical relevance remains largely incomplete. To systematically decipher the effects of human genetic variants, we obtained whole genome sequencing data for 809 individuals from 233 primate species, and identified 4.3 million common protein-altering variants with orthologs in human. We show that these variants can be inferred to have non-deleterious effects in human based on their presence at high allele frequencies in other primate populations. We use this resource to classify 6% of all possible human protein-altering variants as likely benign and impute the pathogenicity of the remaining 94% of variants with deep learning, achieving state-of-the-art accuracy for diagnosing pathogenic variants in patients with genetic diseases.

One Sentence Summary: Deep learning classifier trained on 4.3 million common primate missense variants predicts variant pathogenicity in humans.

Main Text:

A scalable approach for interpreting the effects of human genetic variants and their impact on disease risk is urgently needed to realize the promise of personalized genomic medicine (*1-3*). Out of more than 70 million possible protein-altering variants in the human genome, only ~0.1% are annotated in clinical variant databases such as ClinVar (*4*), with the remainder being variants of uncertain clinical significance (*5, 6*). Despite collaborative efforts by the scientific community, the rarity of most human genetic variants has meant that progress towards deciphering personal genomes has been incremental (*7, 8*). Consequently, clinical sequencing tests frequently return without definitive diagnoses, a frustrating outcome for both patients and clinicians (*9, 10*). In certain cases, patients have needed to be recontacted and diagnoses reversed when the presumed pathogenic variant was later found to be a common variant in previously understudied human populations (*11-13*). Common variants can often be ruled out as the cause of penetrant genetic disease, since their high frequency in the population indicates that they are tolerated by natural selection, aside from rare exceptions due to founder effects and balancing selection (*14-16*).

An emerging strategy for solving clinical variant interpretation on a genome-wide scale is the use of information from closely related primate species to infer the pathogenicity of orthologous human variants (*17*). Because chimpanzees and humans share 99.4% protein sequence identity (*18*), a protein-altering variant present in one species can be expected to produce similar effects on the protein in the other species. By conducting population sequencing studies in closely related non-human primate species, it is feasible to systematically catalog common variants and rule these out as pathogenic in human, analogous to how sequencing more diverse human populations has helped to advance clinical variant interpretation (*8, 17*). Nonetheless, earlier work (*17*) was limited by the very small primate population sequencing datasets available, which bounded the number of common variants discovered, and the scale of machine learning classifiers that could be trained.

RESULTS

A database of 4.3 million benign missense variants across the primate lineage

To expand upon this strategy, we sequenced 703 individuals from 211 primate species (*19*), and aggregated these with data from previous studies (*19-26*), yielding a total of 809 individuals from 233 species. We identified 4.3 million unique missense (protein-altering) variants and 6.7 million unique synonymous (non-protein altering) variants (Fig. 1A), after excluding variants at positions that lacked unambiguous 1:1 mapping with human, or which resulted in non-concordant amino acid translation outcomes because of changes at neighboring nucleotides (fig. S1). The species selected for sequencing represent close to half of the 521 extant primate species on Earth (*27*) and cover all major primate families, from Old World monkeys and New World monkeys to lemurs and tarsiers. We targeted a small number of individuals per species (3.5 on average) to ensure that we primarily sampled common variants that have been filtered by natural selection rather than rare mutations (fig. S2).

Compared to the genome Aggregation Database (gnomAD) cohort of 141,456 human individuals from diverse populations (*28, 29*), the primate sequencing cohort contained ~20% more exome

variants despite sequencing 1/175th the number of individuals (Fig. 1A and fig. S3), attesting to the remarkable genetic diversity present in non-human primate species (*19, 30*), many of which are critically endangered (*31*). The overlap of primate variants with gnomAD was low, consistent with independent mutational origins in each species (fig. S3). Out of the 22 million possible synonymous variants in the human genome, 30% were observed in the primate cohort, compared to just 6% of possible missense mutations (Fig. 1B). Because de novo mutations would have laid down unbiased proportions of missense and synonymous variants, the observed depletion of missense mutations in the primate cohort is consistent with the majority of newly-arising human missense mutations being removed by natural selection due to their deleteriousness (*8, 32-34*). The surviving missense variants are seen at high frequencies in primate populations, and represent a subset of missense variants that have tolerated filtering by natural selection and are unlikely to be pathogenic (*35*).

Missense variants from the primate cohort are strongly enriched for benign consequence in the ClinVar clinical variant database (Fig. 1C). Amongst ClinVar variants with higher review levels (2-star and above, indicating consensus by multiple submitters) (*4*), missense variants found in at least one non-human primate species were Benign or Likely Benign ~99% of the time, compared to 63% for ClinVar missense variants in general, and 80% for missense variants seen in gnomAD (Fig. 1C). The high fraction of pathogenic variants in gnomAD is consistent with the majority of these variants having arisen recently. Indeed, recent exponential human population growth introduced large numbers of rare variants through random de novo mutations (95% of variants in the gnomAD cohort are at < 0.01% population allele frequency), without sufficient time for selection to purge deleterious variants from the population (*36-40*). Consequently, the gnomAD cohort provides a comparatively unfiltered look at variation caused by random mutations, whereas primate common variants represent the subset of random mutations that have survived.

The regions of human disease genes that were most densely populated by ClinVar pathogenic variants were also strongly depleted for primate common variants, with examples shown for *CACNA1A* (Fig. 1D) and *CREBBP* (fig. S4), genes responsible for familial epilepsy (*41, 42*) and Rubinstein-Taybi syndrome (*43, 44*). Missense variants in the gnomAD cohort were partially depleted within these same critical regions (Fig. 1D and fig. S4), indicating that humans and primates experience similar selective pressures. However, deleterious variants were incompletely removed in humans, consistent with the shorter amount of time they were exposed to natural selection.

Prior to using primate data as an indicator of benign consequence in a diagnostic setting, it is vital to understand why a handful of human pathogenic ClinVar variants appear as tolerated common variants in primates. Our clinical laboratory independently reviewed evidence for each of the 36 ClinVar pathogenic variants that appeared in the primate cohort, according to ACMG guidelines (*14*). Among these 36 variants, 8 were reclassified as variants of uncertain significance based on insufficient evidence of pathogenicity in the literature and an additional 9 were hypomorphic or mild clinical variants (table S1). The remaining 19 variants appear to be truly pathogenic in human, and are presumably tolerated in primate because of primate-human differences, such as interactions with changes in the neighboring sequence context (*45, 46*). In one such example, a compensatory synonymous sequence change at an adjacent nucleotide explains why the variant is benign in primate, but creates a pathogenic splice defect in human (Fig. 1E). We also expect that some of the variants identified among primates are rare pathogenic variants by chance, despite the small number of individuals sequenced within each species. By

expanding our cohort to sequence a large number of individuals per species, we would definitively exclude rare variation from our catalogue of primate variation, as well as grow the database of benign variants to improve clinical variant interpretation.

As evolutionary distance from human increases, cases where the surrounding sequence context has changed sufficiently to alter the effect of the variant should also increase, until common variants in more distant species could no longer be reliably counted on as benign in human. We examined variation in each major branch of the primate tree, as well as variation from mammals (mouse, rat, cow, dog), chicken, and zebrafish, and evaluated their pathogenicity in ClinVar (Fig. 1F). Common variants from species throughout the primate lineage, including more distant branches such as lemurs and tarsiers, varied from 98.6% to 99% benign in the human ClinVar database, but this dropped to 87% for placental mammals, and 71% for chicken. The high fraction of variants that are pathogenic in human, yet tolerated as common variants in more distant vertebrates, indicates that selection on orthologous variants diverges substantially in distantly-related species, as a consequence of changes in the surrounding sequence context and other differences in the species' biology (fig. S5).

We have made the primate population variant database, which contains over 4.3 million likely benign missense variants, publicly available at https://primad.basespace.illumina.com as a reference for the genomics community. Overall, this resource is over 50 times larger than ClinVar in terms of number of annotated missense variants, and consists almost entirely of variants of previously unknown significance. Most primate variants are rare or absent in the human population, with 98% of these variants at allele frequency < 0.01% (fig. S6). This makes it challenging to establish their pathogenicity through other means, since even the largest sequencing laboratories would be unlikely to observe any given variant in more than one unrelated patient. Despite their rarity, the subset of human variants that appear in primates have a low missense : synonymous ratio consistent with being depleted of deleterious missense variants (Fig. 1G). This contrasts with the high missense : synonymous ratio for rare human variants in the overall gnomAD cohort, which approaches the 2.2:1 ratio expected for random de novo mutations in the absence of selective constraint (*47*). At higher allele frequencies, natural selection has had more time to purge deleterious missense variants, allowing the human missense : synonymous ratio to start to converge toward the ratio observed for the subset of human variants that are present in other primates.

Gene-level selective constraint in humans versus non-human primates

The primate variant resource makes it possible to compare natural selection acting on individual genes across the primate lineage and identify human-specific evolutionary differences. Since the current primate cohort only contains an average of 3-4 individuals per species, we focused on comparing selective constraint in human genes versus primates as a whole. We found that the missense : synonymous ratios of individual genes were well-correlated between human and primates (Spearman r = 0.637) (Fig. 2A), indicating that genes which were depleted for deleterious missense mutations in human were also consistently depleted throughout the primate lineage. Moreover, the missense : synonymous ratios of both human and primate genes correlated similarly well with the probability of genes being loss of function intolerant (pLI) (Spearman correlation -0.534 and -0.489, respectively) (*28*). Had there been substantial divergence between human and primate, pLI, an independent metric derived from human

protein-truncating variation, would have been expected to show much clearer agreement with human missense : synonymous ratios than primate.

To measure the selective constraint on each gene, we calculated the observed versus expected number of variants per gene, using trinucleotide mutation rates to model the expected probability of observing each variant (fig. S7) (*28, 29*). We modeled each primate species separately to account for differences in genetic diversity and the number of individuals sampled per species. The expected and observed counts of synonymous variants were highly correlated in both the gnomAD and primate cohorts, indicating that our model accurately captured the background distribution of neutral mutations (Fig. 2B; Spearman correlation 0.933 and 0.949, respectively). In contrast, for missense variants the expected and observed counts per gene diverged substantially (Spearman correlation 0.896 and 0.561 for human and primate, respectively), due to depletion of deleterious missense variants by natural selection in highly constrained genes (for example, high pLI genes). The most highly constrained genes were almost completely scrubbed of common missense variants in the primate cohort, whereas rare missense variants in the gnomAD cohort were depleted to a more modest extent due to the large sample size of gnomAD (Fig. 2C).

We next aimed to identify genes whose selective constraint was different in human compared to the rest of the primate lineage, a task made difficult by differences in diversity, allele frequency, and sample size between the human and primate cohorts (*34, 48, 49*). To this end, we developed two orthogonal strategies, and took the intersection of genes identified under both approaches. First, we used population genetic modeling (*34, 50, 51*) to estimate the average selection coefficient, s, ranging from 0 (benign) to 1 (severely pathogenic), of missense mutations in each gene, using a model of recent human population growth (figs. S7 and S8). We fit a single value of s per gene across non-human primate species, and identified genes that differed between $s_{primate}$ and s_{human} using a likelihood ratio test, which we validated using population simulations (fig. S9). In a second approach, we fit a curve approximating the relationship between human and primate missense : synonymous ratios using a Poisson generalized linear mixed model (*52*), and identified genes where the observed human missense : synonymous ratio deviated from what would have been expected given the gene's missense : synonymous ratio in primates (fig. S10). We also adjusted for gene length to account for shorter genes having more variability in their missense : synonymous ratio measurements than longer genes. The two methods were broadly concordant, with a Spearman correlation of 0.80 between the genes' effect sizes in the two tests. Estimates of selection coefficients and observed and expected counts for each gene in human and primate are provided in table S2.

In total, we found 39 genes where selective constraint differed significantly between human and other primates under both methods (Benjamini-Hochberg FDR < 0.05 (*53*); Fig. 2D). The top three genes where s_{human} decreased the most relative to $s_{primate}$ were *CFTR*, *GJB2*, and *CD36*, autosomal recessive disease genes for cystic fibrosis (*54*), hereditary deafness (*55*), and platelet glycoprotein deficiency (*56*), respectively. All three genes are known for deleterious mutations that are unusually common in local geographic human populations (*57-60*), suggesting that they may be experiencing reduced selection due to heterozygote advantage that protects against specific environmental pathogens (*60-64*). On the other end of the spectrum, *TERT*, known for its role in maintaining telomere length (*65, 66*), was among the top genes where s_{human} increased the most relative to $s_{primate}$. Humans have adapted to a much longer lifespan compared with other primate species, which have a median lifespan of 20-30 years, suggesting that increased selection

on *TERT* may have occurred as part of human adaption towards extended longevity. We note that with the current size of the primate cohort, it is not possible to distinguish whether the increased selection on *TERT* occurred only in humans, or if it is part of a gradual trend towards extended longevity that began earlier in the great ape lineage, which also have longer lifespans relative to other primates (~40 years). Expanding the primate cohort by sequencing more individuals per species would improve detection of additional species-specific and lineage-specific evolutionary adaptation, and shed light on the evolutionary path that led to the present human condition.

PrimateAI-3D, a deep learning network for classifying protein-altering variants

We constructed PrimateAI-3D, a semi-supervised 3D-convolutional neural network for variant pathogenicity prediction, which we trained using 4.5 million common missense variants with likely benign consequence (Fig. 3A). In a departure from prior deep learning architectures that operated on linear sequence (*17, 67*), we voxelized the 3D structure of the protein at 2 Angstrom resolution (figs. S11 and S12) and used 3D-convolutions to enable the network to recognize key structural regions that may not be apparent from sequence alone (Fig. 3A). As an example, we show PrimateAI-3D predictions for *STK11* (Fig. 3B), the tumor suppressor gene responsible for Peutz-Jeghers hereditary polyposis syndrome (*68-71*), with each amino acid position colored by the average PrimateAI-3D score at that position. Common primate variants used for training and annotated ClinVar pathogenic variants from separate parts of the linear sequence form distinct clusters in 3D space. Although ClinVar variants are shown for illustration, it is important to note that the network was not trained on either human-engineered features or annotated variants from clinical variant databases, thereby avoiding potential human biases in variant annotation. Rather, it learns to infer pathogenicity based on the local enrichment or depletion of common primate variants, taking only the protein's multiple sequence alignment and 3D structure as inputs.

PrimateAI-3D can utilize protein structures from either experimental sources or computational prediction (*72-76*); we used AlphaFold DB (*72, 73*) and HHpred (*74*) predicted structures for the broadest coverage across human genes. For training data, we incorporated all common missense variants from the 233 non-human primate species (*17*), and common human missense variants (allele frequency > 0.1% across populations) in gnomAD (*28, 29*), TOPMed (*77, 78*), and UK Biobank (*79, 80*), resulting in a total of 4.5 million unique missense variants of likely benign consequence. This dataset covers 6.34% of all possible human missense variants, and is over 50-fold larger than the current ClinVar database (79,381 missense variants after excluding variants of uncertain significance and those with conflicting annotations), greatly enlarging the training dataset available for machine learning approaches. Because the training dataset consists only of variants labeled as benign, we created a control set of randomly selected variants that were matched to the common variants by trinucleotide mutation rate, and trained PrimateAI-3D to separate common variants from matched controls as a semi-supervised learning task.

In parallel with the variant classification task, we generated amino acid substitution probabilities for each position in the protein by masking the residue and using the sequence context to predict the missing amino acid, borrowing from language model architectures that are trained to predict missing words in sentences (*81, 82*). We trained both a 3D convolutional "fill-in-the-blank" model, which tasked the network with predicting the missing amino acid in a gap in the voxelized 3D protein structure, and separately, a language model utilizing the transformer

architecture to predict the missing amino acid using the surrounding multiple sequence alignment as context (*83*). We implemented these models as additional loss functions to further refine the PrimateAI-3D predictions (fig. S13). We also trained a variational autoencoder (*67*) on multiple sequence alignments and found that it performed comparably to our transformer architecture (fig. S14). Hence, we incorporated the average of their predictions in the loss function, which performed better than either alone.

We evaluated PrimateAI-3D and 15 other published machine learning methods (*67, 84*) on their ability to distinguish between benign and pathogenic variants along six different axes (Fig. 3C, 3D, and fig. S15): predicting the effects of rare missense variants on quantitative clinical phenotypes in a cohort of 200,643 individuals from the UK Biobank (UKBB); distinguishing missense de novo mutations (DNM) seen in 31,058 patients with neurodevelopmental disorders (*85-87*) (DDD) from de novo missense mutations in 2,555 healthy controls (*88-93*); distinguishing de novo missense mutations seen in 4,295 patients with autism spectrum disorders (*88-94*) (ASD) from de novo missense mutations in the shared set of 2,555 healthy controls; distinguishing de novo missense mutations seen in 2,871 patients with congenital heart disease (*95*) (CHD) from de novo missense mutations in the shared set of 2,555 healthy controls; separating annotated ClinVar benign and pathogenic variants (ClinVar) (*4*); and average correlation with in vitro deep mutational scan experimental assays across 9 genes (*96-105*) (DMS assays). Our set of clinical benchmarks is the most comprehensive to date, and has a particular focus on rigorously testing the performance of classifiers on large patient cohorts across a diverse range of real world clinical settings (table S3).

For the UK Biobank benchmark, we analyzed 200,643 individuals with both exome sequencing data and broad clinical phenotyping, and identified 42 genes where the presence of rare missense variants was associated with changes in a quantitative clinical phenotype controlling for confounders such as population stratification, age, sex, and medications (table S4). These gene-phenotype associations included diverse clinical lab measurements such as low-density lipoprotein (LDL) cholesterol (increased by rare missense variants in *LDLR*, decreased by variants in *PCSK9*), blood glucose (increased by variants in *GCK*), and platelet count (increased by variants in *JAK2*, decreased by variants in *GP1BB*), as well as other quantitative phenotypes such as standing height (increased by variants in *ZFAT*) (table S4). To test each classifier's ability to distinguish between pathogenic and benign missense variants, we measured the correlation between pathogenicity prediction score and quantitative phenotype for patients carrying rare missense variants in each of these genes. We report the average correlation across all gene-phenotype pairs for each classifier, taking the absolute value of the correlation, since these genes may be associated with either increase or decrease in the quantitative clinical phenotype.

The neurodevelopmental disorders cohort (DDD), autism spectrum disorders cohort (ASD), and congenital heart disease cohort (CHD) are among the largest published trio-sequencing studies to date, and consist of thousands of families with a child with rare genetic disease and their unaffected parents. In each cohort, we cataloged de novo missense mutations that appeared in affected probands but were absent in their parents, as well as de novo missense mutations that appeared in a set of shared healthy controls. We evaluated the ability of each classifier to separate the de novo missense mutations that appear in cases versus controls on the basis of their prediction scores, using the Mann-Whitney U test to measure performance.

PrimateAI-3D outperformed all other classifiers at distinguishing pathogenic from benign variants in the four patient cohorts we tested (UKBB, DDD, ASD, CHD); it was also the top performer at separating pathogenic from benign variants in the ClinVar annotation database, and had the highest average correlation with the deep mutational scan assays (Fig. 3D and fig. S15). After PrimateAI-3D, there was no clear runner-up, with second place occupied by six different classifiers in the six different benchmarks. We observed a moderate correlation between the performance of different classifiers in UKBB and DDD (Spearman r = 0.556; Fig. 3C), which are the two largest clinical cohorts and therefore likely the most robust for benchmarking (with 200,643 and 33,613 patients, respectively), but outside of PrimateAI-3D, strong performance of a classifier on one task had limited generalizability to other tasks. Our results underscore the importance of validating machine learning classifiers along multiple dimensions, particularly in large real-world cohorts, to avoid overgeneralizing a classifier's performance based upon an impressive showing along a single axis.

PrimateAI-3D's top-ranked performance at separating benign and pathogenic missense variants in ClinVar was unexpected, since the other machine learning classifiers (with the exception of EVE) were trained either directly on ClinVar, or on other variant annotation databases with a high degree of content overlap. Because they are primarily based on variants described in the literature, clinical variant databases are subject to ascertainment bias (*12, 106, 107*), which may have contributed to supervised classifiers picking up on tendencies of human variant annotation that are unrelated to the task of separating benign from pathogenic variants (figs. S16, S17, and S18). Given the challenges with human annotation, we also investigated whether PrimateAI-3D could assist in revising incorrectly labeled ClinVar variants, by comparing annotations in the current ClinVar database and those from a September 2017 snapshot. Disagreement between PrimateAI-3D and the 2017 version of ClinVar was highly predictive of future revision and the odds of revision increased with PrimateAI-3D confidence (fig. S19). Among variants with the 10% most confident PrimateAI-3D predictions, the odds of revision were 10-fold elevated if PrimateAI-3D was in disagreement with the ClinVar label ($P < 10^{-14}$).

The performance of PrimateAI-3D on clinical variant benchmarks scaled directly with training dataset size, indicating that additional primate sequencing data will be the key to unlocking further gains (Fig. 4 and fig. S20). The current primate cohort already covers 30% of all possible synonymous variants in the human genome, despite containing only 809 individuals from 233 species (Fig. 4B). By increasing the number of species and the number of individuals sequenced per species, we expect to saturate the majority of the remaining tolerated substitutions in the human genome (fig. S21), including both coding and noncoding variation, leaving the remaining deleterious variants to be deduced by process of elimination.

Discovery of candidate disease genes for neurodevelopmental disorders

We applied PrimateAI-3D to improve statistical power for discovering candidate disease genes that are enriched for pathogenic de novo mutations in the neurodevelopmental disorders cohort (fig. S22). De novo missense mutations from affected individuals in the DDD cohort (*87*) were enriched 1.36-fold above expectation, based on estimates of background mutation rate using trinucleotide context (*47*). We selected a PrimateAI-3D classification threshold of 0.821, which called an equal number of pathogenic missense mutations (n=7,238) as the excess of de novo missense mutations in the cohort (Fig. 5A). Stratifying missense mutations by this threshold

increased enrichment of pathogenic de novo missense mutations to 2.0-fold, substantially increasing statistical power for disease gene discovery in the cohort (Fig. 5B).

By applying PrimateAI-3D to prioritize pathogenic missense variants, we identified 290 genes associated with intellectual disability at genome-wide significance ($P < 6.4 \times 10^{-7}$) (Table 1), of which 272 were previously discovered genes that either appeared in the Genomics England intellectual disability gene panel (*108*), or were already identified in the prior study (*109*) without stratifying missense variants (table S5). We excluded two genes, *BMPR2* and *RYR1* as borderline significant genes that already had well-annotated non-neurological phenotypes. Further clinical studies are needed to independently validate this list of candidate genes and understand their range of phenotypic effects.

Discussion

Our results demonstrate the successful pairing of primate population sequencing with state-of-the-art deep learning models to make meaningful progress towards solving variants of uncertain significance. Primate population sequencing and large-scale human sequencing are likely to fill complementary roles in advancing clinical understanding of human genetic variants. From the perspective of acquiring additional benign variants to train PrimateAI-3D, humans are not suitable, as the discovery of common human variants (>0.1% allele frequency) plateaus at roughly ~100,000 missense variants after only a few hundred individuals (*17*), and further population sequencing into the millions mainly contributes rare variants which cannot be ruled out for deleterious consequence. On the other hand, these rare human variants, because they have not been thoroughly filtered by natural selection, preserve the potential to exert highly penetrant phenotypic effects, making them indispensable for discovering new gene-phenotype relationships in large population sequencing and biobank studies. Fittingly, classifiers trained on common primate variants may accelerate these target discovery efforts, by helping to differentiate between benign and pathogenic rare variation.

The genetic diversity found in the 520 known non-human primate species is the result of ongoing natural experiments on genetic variation that have been running uninterrupted for millions of years. Today, over 60% of primate species on Earth are threatened with extinction in the next decade due to man-made factors (*31*). We must decide whether to act now to preserve these irreplaceable species, which act as a mirror for understanding our genomes and ourselves, and are each valuable in their own right, or bear witness to the conclusion of many of these experiments.

Materials and Methods

Primate polymorphism data

We aggregated high-coverage whole genomes of 809 primate individuals across 233 primate species, including 703 newly sequenced samples and 106 previously sequenced samples from the Great Ape Genome project (*19*). Samples that passed quality evaluation were then aligned to 32 high-quality primate reference genomes (*110*) and mapped to the GRCh38 human genome build.

We developed a random forest (RF) classifier to identify false positive variant calls and errors resulting from ambiguity in the species mapping. In addition, we removed variants that fell in primate codons that did not match the human codon at that position, as well as those residing in primate transcripts with likely annotation errors. We also devised quality metrics based on the distribution of RF scores and Hardy-Weinberg equilibrium, and developed a unique mapping filter to exclude variants in regions of non-unique mapping between primate species.

Identifying differential selection between humans and primates via population modeling

We first established a neutral background distribution of mutation rates per gene for each primate species by fitting the Poisson Random Field (PRF) model to the segregating synonymous variants in each species. The observed number of segregating synonymous sites is a Poisson random variable, with the mean determined by mutation rate, demography, and sample size (*34*). For simplicity, we assumed an equilibrium (i.e. constant) demography for all species besides human; for human, we used Moments (*51*) to find a best fitting demographic history based on the folded site frequency spectrum of synonymous sites. We adopted a Gamma distributed prior on mutation rates, which also accounts for the impact of GC content on mutation rate. We optimized the prior parameters via maximum likelihood and computed the posterior distribution of the mutation rate per gene.

The number of segregating nonsynonymous sites is modeled as a Poisson random variable similar to synonymous sites with additional selection parameters. We assumed that every nonsynonymous mutation in a gene shares the same population scaled selection coefficient γ_{ig}. To explicitly estimate selection coefficient of each gene per species, we devised a two-step procedure analogous to an EM algorithm to control for differences in population size across species.

To identify genes where human constraint is different from non-human primate selection, we developed a likelihood ratio test to test whether population scaled selection coefficients are significantly different between human and other primates. We then assessed whether our population genetic modeling improved the correlation of selection estimates of our primate data with previous gene-constraint metrics in humans, including pLI (*28*) and s_het (*111*). To validate the performance of our model, we performed population genetic simulations.

Poisson generalized linear mixed modeling of selection between humans and primates

In addition to population genetics model described above, we also applied an orthogonal approach to detect differences in selection between humans and primates based on missense-to-synonymous ratio (MSR). We fit a Poisson generalized linear mixed model (GLMM) to the pooled polymorphic synonymous and missense mutations across all primates to estimate the depletion of missense variants in each gene. Then, we fit a second Poisson GLMM to the human

data, controlling for the primate depletion estimates, and compared the pooled primate MSR to the human MSR for each gene.

PrimateAI-3D Model

PrimateAI-3D is a 3D convolutional neural network that uses protein structures and multiple sequence alignments (MSA) to predict the pathogenicity of human missense variants. To generate the input for a 3D convolutional neural network, we voxelized the protein structure and evolutionary conservation in the region surrounding the missense variant. The network was trained to optimize three objectives: distinction between benign and unknown human variants; prediction of a masked amino acid at the variant site; per-gene variant ranks based on protein language models.

Protein structures and multiple sequence alignments

For 341 species, we used vertebrate and mammal MSAs from UCSC Multiz100 (*112, 113*) and Zoonomia (*114*). Another 251 species appeared in Uniprot for least 75% of all human proteins (*115*). For each protein, alignments from all 341+251=592 species were merged. Human protein structures were taken from AlphaFold DB (June 2021) (*73*). Proteins that did not sequence-match exactly to our hg38 proteins (2590; 13.5%) were homology modeled using HHpred (*74*) and Modeller (*116*).

Protein voxelization and voxel features

A regular sized 3D grid of 7x7x7 voxels, each spanning 2Åx2Åx2Å, was centered at the Cα atom of the residue containing the target variant (Fig. S11). For each voxel, we provided a vector of distances between its center and the nearest Cα and Cβ atoms of each amino acid type (Fig. S11; details in Supplementary Text section 1). We also provided additional voxel features including the pLDDT confidence metric from AlphaFold DB (Fig. S12), and the evolutionary profile, consisting of each amino acid's frequency at the corresponding position in the 592 species alignment.

Model architecture

The first layers of the PrimateAI-3D model reduce the voxel tensor to a 64-vector through repeated valid-padded 3D convolutions with a kernel size of 3x3x3. A final hidden dense layer transforms this 64-length vector into a 20-length vector, corresponding to one output unit per amino acid at that position. The model was trained simultaneously using multiple loss functions to optimize the following complementary aspects of pathogenicity:

<u>Benign primate variants</u>

Using 4.5 million benign missense variants from primates, we sampled the same number of unknown variants from the set of all possible human missense variants, with the distribution of mutational probabilities matching the benign set, based on a trinucleotide mutation rate model. Variants for the same protein position were combined in a 20-length vector (benign: 0, unknown: 1) which was the target label for the network. We used mean squared error (MSE) as the loss function for non-missing labels and ignored missing labels.

<u>3D fill-in-the-blank</u>

We removed all atoms of a target residue before voxelization, discarding any information about the residue from the input tensor to the network. The network was then trained to predict a 20-

length vector, labeled 0 (benign) for amino acids that occur at the target site in any of the 592 species and 1 (pathogenic) otherwise. All human protein positions with at least one possible missense variant were included in this dataset.

Variant ranks from language models

For each gene, we took the average pathogenicity ranking from two protein language models, PrimateAI language model (PrimateAI LM, described below) and our reimplementation of the EVE variational autoencoder algorithm which we extended to all human proteins (EVE*) (*67*). We calculated the pairwise logistic rank loss as described in Pasumarthi *et al.*(*117*).

PrimateAI Language Model

The PrimateAI language model (PrimateAI LM) is a MSA transformer (*83*) for fill-in-the-blank residue classification, which was trained end-to-end on MSAs of UniRef-50 proteins (*118, 119*) to minimize an unsupervised masked language modelling (MLM) objective (*81*). Our model requires ~50x less computation for training than previous MSA transformers due to several improvements in architecture and training (Fig. S9).

Model training procedure

Each batch had the same number of samples from each of the three variant datasets (~33 with a batch size of 100). For the language model ranks dataset, all 33 samples had to come from the same protein. The number of times a protein was chosen for a batch was proportional to the length of the protein. In order to make our model robust against protein orientations, we randomly rotated the protein atomic coordinates in 3D before voxelizing a variant.

Model Evaluation

We compared performance of our model and other models (*84*) on variants for which all models had scores. Deep mutational scanning assays were available for 9 human genes: Amyloid-beta (*102*), *YAP1* (*96*), *MSH2* (*120*), *SYUA* (*101*), *VKOR1* (*121*), *PTEN* (*99, 100*), *BRCA1* (*122*), *TP53* (*123*), and *ADRB2* (*124*). For each assay and prediction model, we calculated the absolute Spearman rank correlation between prediction and assay scores. The UK Biobank dataset (*79, 80*) contains 42 gene-phenotype pairs which were significantly associated by rare variant burden testing using all rare missense variants, without applying missense pathogenicity prioritization. The evaluation was the same as with DMS assays, except that correlations were calculated from the quantitative phenotypes of individuals carrying the variant, instead of the assay score for the variant. For ClinVar (*4*), we filtered to high-quality 2-star variants and evaluated model performance by calculating per-gene area under the receiver operating characteristic curve (AUC). For the rare disease cohorts, we collected de novo missense mutations from patients with developmental disorders (*85-87*), autism spectrum disorders (*88-94*) or congenital heart disorders (*95*). For all three datasets, we compared against DNMs from healthy controls (*88-93*). We applied the Mann-Whitney U test to measure how well each model's prediction scores could distinguish patient variants from control variants.

References and Notes:

1. D. G. MacArthur *et al.*, Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476 (2014).

2. R. L. Nussbaum, H. L. Rehm, ClinGen, ClinGen and Genetic Testing. *N Engl J Med* **373**, 1379 (2015).
3. H. L. Rehm *et al.*, ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**, 2235-2242 (2015).
4. M. J. Landrum *et al.*, ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868 (2016).
5. X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241 (2016).
6. P. D. Stenson *et al.*, The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1-9 (2014).
7. H. L. Rehm, Evolving health care through personal genomics. . *Nat Rev Genet* **18**, 259-267 (2017).
8. N. Whiffin *et al.*, Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* **19**, 1151-1158 (2017).
9. S. Caspar *et al.*, Clinical sequencing: from raw data to diagnosis with lifetime value. *Clinical genetics* **93**, 508-519 (2018).
10. Y. Yang *et al.*, Molecular findings among patients referred for clinical whole-exome sequencing. *Jama* **312**, 1870-1879 (2014).
11. J. A. SoRelle, D. M. Thodeson, S. Arnold, G. Gotway, J. Y. Park, Clinical Utility of Reinterpreting Previously Reported Genomic Epilepsy Test Results for Pediatric Patients. *JAMA Pediatr* **173**, e182302 (2019).
12. N. Shah *et al.*, Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *Am J Hum Genet* **102**, 609-619 (2018).
13. O. Campuzano *et al.*, Reanalysis and reclassification of rare genetic variants associated with inherited arrhythmogenic syndromes. *EBioMedicine* **54**, 102732 (2020).
14. S. Richards *et al.*, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424 (2015).
15. Y. E. Kim, C. S. Ki, M. A. Jang, Challenges and Considerations in Sequence Variant Interpretation for Mendelian Disorders. *Ann Lab Med* **39**, 421-429 (2019).
16. M. Slatkin, A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. *The American Journal of Human Genetics* **75**, 282-293 (2004).
17. L. Sundaram *et al.*, Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).
18. C. S. A. Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
19. J. Prado-Martinez *et al.*, Great ape genome diversity and population history. *Nature* **499**, 471-475 (2013).
20. Z. Fan *et al.*, Ancient hybridization and admixture in macaques (genus Macaca) inferred from whole genome sequences. *Mol Phylogenet Evol* **127**, 376-386 (2018).
21. Z. Liu *et al.*, Genomic Mechanisms of Physiological and Morphological Adaptations of Limestone Langurs to Karst Habitats. *Mol Biol Evol* **37**, 952-968 (2020).
22. L. Wang *et al.*, A high-quality genome assembly for the endangered golden snub-nosed monkey (Rhinopithecus roxellana). *Gigascience* **8**, (2019).
23. C. Zoonomia, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240-245 (2020).

24. B. J. Evans *et al.*, Speciation over the edge: gene flow among non-human primate species across a formidable biogeographic barrier. *R Soc Open Sci.* **4**, 170351 (2017).
25. L. Yu *et al.*, Genomic analysis of snub-nosed monkeys (Rhinopithecus) identifies genes and processes related to high-altitude adaptation. *Nat Genet* **48**, 947-952 (2016).
26. N. Osada, K. Matsudaira, Y. Hamada, S. Malaivijitnond, Testing sex-biased admixture origin of macaque species using autosomal and X-chromosomal genomic sequences. *Genome Biol. Evol.* **13**, (2021).
27. A. B. Rylands, R. A. Mittermeier, *Primate Behavioral Ecology*. (Routledge, New York, ed. 6, 2021), pp. 407–428.
28. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
29. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
30. E. M. Leffler *et al.*, Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**, e1001388 (2012).
31. A. Estrada *et al.*, Impending extinction crisis of the world's primates: Why primates matter. *Sci Adv* **3**, e1600946 (2017).
32. T. Ohta, Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-98 (1973).
33. D. E. Reich, E. S. Lander, On the allelic spectrum of human disease. *Trends Genet* **17**, 502-510 (2001).
34. S. A. Sawyer, D. L. Hartl, Population genetics of polymorphism and divergence. *Genetics* **132**, 1161-1176 (1992).
35. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610-618 (2007).
36. W. Fu *et al.*, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).
37. Y. B. Simons, M. C. Turchin, J. K. Pritchard, G. Sella, The deleterious mutation load is insensitive to recent population history. *Nature genetics* **46**, 220-224 (2014).
38. R. Do *et al.*, No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics* **47**, 126-131 (2015).
39. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology* **18**, e3000586 (2020).
40. I. Mathieson, G. McVean, Demography and the age of rare variants. *PLoS Genet* **10**, e1004528 (2014).
41. L. Damaj *et al.*, CACNA1A haploinsufficiency causes cognitive impairment, autism and epileptic encephalopathy with mild cerebellar symptoms. *Eur J Hum Genet* **23**, 1505-1512 (2015).
42. K. Reinson *et al.*, Biallelic CACNA1A mutations cause early onset epileptic encephalopathy with progressive cerebral, cerebellar, and optic nerve atrophy. *Am J Med Genet A* **170**, 2173-2176 (2016).
43. A. Bentivegna *et al.*, Rubinstein-Taybi Syndrome: spectrum of CREBBP mutations in Italian patients. *BMC Med Genet* **7**, 77 (2006).
44. M. Stef *et al.*, Spectrum of CREBBP gene dosage anomalies in Rubinstein-Taybi syndrome patients. *Eur J Hum Genet* **15**, 843-847 (2007).
45. A. S. Kondrashov, S. Sunyaev, F. A. Kondrashov, Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* **99**, 14878-14883 (2002).
46. D. M. Jordan *et al.*, Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524**, 225-229 (2015).

47. K. E. Samocha *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).
48. C. D. Bustamante, J. Wakeley, S. Sawyer, D. L. Hartl, Directional selection and the site-frequency spectrum. *Genetics* **159**, 1779-1788 (2001).
49. X. Huang *et al.*, Inferring genome-wide correlations of mutation fitness effects between populations. *Molecular Biology and Evolution*.
50. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (2009).
51. J. Jouganous, W. Long, A. P. Ragsdale, S. Gravel, Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics* **206**, 1549-1567 (2017).
52. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,* **67**, 1-48 (2015).
53. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57**, 289-300 (1995).
54. R. K. Rowntree, A. Harris, The phenotypic consequences of CFTR mutations. *Annals of human genetics* **67**, 471-485 (2003).
55. S. A. Wilcox *et al.*, High frequency hearing loss correlated with mutations in the GJB2 gene. *Human genetics* **106**, 399-405 (2000).
56. H. Shu *et al.*, The role of CD36 in cardiovascular disease. *Cardiovascular Research*, (2020).
57. J. L. Bobadilla, M. Macek Jr, J. P. Fine, P. M. Farrell, Cystic fibrosis: a worldwide analysis of CFTR mutations—correlation with incidence data and application to screening. *Human mutation* **19**, 575-606 (2002).
58. M. H. Chaleshtori *et al.*, High carrier frequency of the GJB2 mutation (35delG) in the north of Iran. *International journal of pediatric otorhinolaryngology* **71**, 863-867 (2007).
59. J. Liu *et al.*, Distribution of CD36 deficiency in different Chinese ethnic groups. *Human Immunology* **81**, 366-371 (2020).
60. T. J. Aitman *et al.*, Malaria susceptibility and CD36 mutation. *Nature* **405**, 1015-1016 (2000).
61. J. E. Common, W.-L. Di, D. Davies, D. P. Kelsell, Further evidence for heterozygote advantage of GJB2 deafness mutations: a link with cell survival. *Journal of medical genetics* **41**, 573-575 (2004).
62. P. D'Adamo *et al.*, Does epidermal thickening explain GJB2 high carrier frequency and heterozygote advantage? *European Journal of Human Genetics* **17**, 284-286 (2009).
63. S. A. Schroeder, D. M. Gaughan, M. Swift, Protection against bronchial asthma by CFTR ΔF508 mutation: a heterozygote advantage in cystic fibrosis. *Nature medicine* **1**, 703-705 (1995).
64. G. B. Pier *et al.*, Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature* **393**, 79-82 (1998).
65. S. E. Bojesen *et al.*, Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics* **45**, 371-384 (2013).
66. B. Heidenreich, R. Kumar, TERT promoter mutations in telomere biology. *Mutation Research/Reviews in Mutation Research* **771**, 15-31 (2017).

67. J. Frazer *et al.*, Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).
68. H. D. Chae, C. H. Jeon, Peutz-Jeghers syndrome with germline mutation of STK11. *Ann Surg Treat Res* **86**, 325-330 (2014).
69. I. Hernan *et al.*, De novo germline mutation in the serine-threonine kinase STK11/LKB1 gene associated with Peutz-Jeghers syndrome. *Clin Genet* **66**, 58-62 (2004).
70. C. Nakanishi *et al.*, Germline mutation of the LKB1/STK11 gene with loss of the normal allele in an aggressive breast cancer of Peutz-Jeghers syndrome. *Oncology* **67**, 476-479 (2004).
71. H. R. Yang, J. S. Ko, J. K. Seo, Germline mutation analysis of STK11 gene using direct sequencing and multiplex ligation-dependent probe amplification assay in Korean children with Peutz-Jeghers syndrome. *Dig Dis Sci* **55**, 3458-3465 (2010).
72. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
73. M. Varadi *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, (2021).
74. J. Söding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**, W244-W248 (2005).
75. M. Kallberg *et al.*, Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* **7**, 1511-1522 (2012).
76. S. Wang, W. Li, S. Liu, J. Xu, RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res* **44**, W430-435 (2016).
77. D. J. Burgess, The TOPMed genomic resource for human health. *Nat Rev Genet* **22**, 200 (2021).
78. D. Taliun *et al.*, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
79. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
80. C. Sudlow *et al.*, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
81. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **1**, 4171-4186 (2019).
82. Y. You *et al.*, in *International Conference on Learning Representations*. (2020).
83. R. M. Rao *et al.*, MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning* **139**, 8844–8856 (2021).
84. X. Liu, C. Li, C. Mou, Y. Dong, Y. Tu, dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12**, 103 (2020).
85. Deciphering Developmental Disorders Study, Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228 (2015).
86. Deciphering Developmental Disorders Study, Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
87. J. Kaplanis *et al.*, Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757-762 (2020).
88. J. Y. An *et al.*, Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, (2018).

89. S. De Rubeis *et al.*, Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
90. I. Iossifov *et al.*, The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
91. I. Iossifov *et al.*, De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).
92. S. J. Sanders *et al.*, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015).
93. S. J. Sanders *et al.*, De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).
94. B. J. O'Roak *et al.*, Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).
95. S. C. Jin *et al.*, Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* **49**, 1593-1601 (2017).
96. C. L. Araya *et al.*, A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences* **109**, 16858–16863 (2012).
97. M. A. Chiasson *et al.*, Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* **9**, (2020).
98. X. Jia *et al.*, Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *American journal of human genetics* **108**, 163-175 (2021).
99. K. A. Matreyek *et al.*, Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50**, 874-882 (2018).
100. T. L. Mighell, S. Evans-Dutson, B. J. O'Roak, A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *American journal of human genetics* **102**, 943-955 (2018).
101. R. W. Newberry, J. T. Leong, E. D. Chow, M. Kampmann, W. F. DeGrado, Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nature Chemical Biology* **16**, 653-659 (2020).
102. M. Seuma, A. J. Faure, M. Badia, B. Lehner, B. Bolognesi, The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *Elife* **10**, e63364 (2021).
103. A. O. Giacomelli *et al.*, Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature genetics* **50**, 1381-1387 (2018).
104. L. M. Starita *et al.*, Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413-422 (2015).
105. E. M. Jones *et al.*, Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *eLife* **9**, e54895 (2020).
106. C. E. G. Amorim *et al.*, The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genet* **13**, e1006915 (2017).
107. B. Quintans, A. Ordonez-Ugalde, P. Cacheiro, A. Carracedo, M. J. Sobrido, Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Appl Transl Genom* **3**, 60-67 (2014).
108. A. R. Martin *et al.*, PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* **51**, 1560-1565 (2019).
109. A. Thormann *et al.*, Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun* **10**, 2373 (2019).

110. L. F. Kuderna *et al.*, A global catalog of whole-genome diversity from 233 primate species *Submitted*.
111. C. A. Cassa *et al.*, Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet* **49**, 806-810 (2017).
112. C. Tyner *et al.*, The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**, D626-D634 (2017).
113. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
114. D. P. Genereux *et al.*, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240-245 (2020).
115. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932 (2015).
116. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815 (1993).
117. R. K. Pasumarthi *et al.*, TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2970--2978 (2019).
118. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288 (2007).
119. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).
120. X. Jia *et al.*, Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet* **108**, 163-175 (2021).
121. M. A. Chiasson *et al.*, Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**, (2020).
122. L. M. Starita *et al.*, Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413-422 (2015).
123. A. O. Giacomelli *et al.*, Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature Genetics* **50**, 1381-1387 (2018).
124. E. M. Jones *et al.*, Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *Elife* **9**, e54895 (2020).
125. S. Gudmundsson *et al.*, Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **597**, E3-E4 (2021).
126. D. Vanderpool *et al.*, Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol* **18**, e3000954 (2020).
127. P. J. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
128. M. A. Eberle *et al.*, A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27**, 157-164 (2017).
129. M. de Manuel *et al.*, Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477-481 (2016).
130. E. M. Leffler, Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J.D. Wall, G. Sella, P. Donnelly, Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578-1582 (2013).

131. K. E. Eilertson, J. G. Booth, C. D. Bustamante, SnIPRE: Selection Inference Using a Poisson Random Effects Model. *PLoS Comput Biol* **8**, e1002806 (2012).
132. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
133. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
134. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
135. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
136. D. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, (2014).
137. M. Mirdita *et al.*, Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170-D176 (2017).
138. M. Steinegger *et al.*, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
139. R. M. Rao *et al.,* M. Meila, T. Zhang, Eds. (PMLR, 2021), vol. 139, pp. 8844-8856.
140. J. L. Ba, J. R. Kiros, G. E. Hinton, paper presented at the Advances in NIPS 2016 Deep Learning Symposium, 2016 2016.
141. D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, (2020).
142. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929-1958 (2014).
143. P. Micikevicius *et al.*, Mixed Precision Training. *International Conference on Learning Representations*, (2018).
144. S. Rajbhandari, J. Rasley, O. Ruwase, Y. He, ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1-16 (2020).
145. P. Bandaru *et al.*, Deconstruction of the Ras switching cycle through saturation mutagenesis. *Elife* **6**, (2017).
146. J. Weile *et al.*, A framework for exhaustively mapping functional missense variants. *Mol Syst Biol* **13**, 957 (2017).
147. L. Brenan *et al.*, Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep* **17**, 1171-1183 (2016).
148. M. M. Awad *et al.*, Acquired Resistance to KRASG12C Inhibition in Cancer. *New England Journal of Medicine* **384**, 2382-2393 (2021).
149. L. Zhang *et al.*, SLCO1B1: Application and Limitations of Deep Mutational Scanning for Genomic Missense Variant Function. *Drug Metab Dispos*, DMD-AR-2020-000264 (2021).
150. V. E. Gray *et al.*, Elucidating the Molecular Determinants of Aβ Aggregation with Deep Mutational Scanning. *G3 (Bethesda)* **9**, 3683-3689 (2019).
151. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248-249 (2010).
152. B.-J. Feng, PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation* **38**, 243-251 (2017).

153. P. Rentzsch, M. Schubach, J. Shendure, M. Kircher, CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine* **13**, 31 (2021).
154. D. Quang, Y. Chen, X. Xie, DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763 (2015).
155. D. Raimondi *et al.*, DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Research* **45**, W201-W206 (2017).
156. N. Malhis, M. Jacobson, S. J. M. Jones, J. Gsponer, LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Research* **48**, W154-W161 (2020).
157. K. A. Jagadeesh *et al.*, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* **48**, 1581-1586 (2016).
158. R. Steinhaus *et al.*, MutationTaster2021. *Nucleic Acids Research* **49**, W446-W451 (2021).
159. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
160. N. M. Ioannidis *et al.*, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).
161. N.-L. Sim *et al.*, SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* **40**, W452-457 (2012).
162. H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, R. Karchin, Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
163. H. A. Shihab *et al.*, An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536-1543 (2015).
164. J. Meier *et al.*, Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, (2021).
165. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15**, 816-822 (2018).
166. P. D. Stenson *et al.*, Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* **45**, 124-126 (2008).
167. P. D. Stenson *et al.*, The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139**, 1197-1207 (2020).
168. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710 (2004).
169. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* **87**, 012707 (2013).

Acknowledgments: We would like to thank Daniel MacArthur, Yun Song, and Mark Daly for helpful discussions, and the gnomAD team at the Broad Institute for their assistance with the website.

Funding: LFKK was supported by an EMBO STF 8286 (to LFKK). MCJ was supported by (NERC) NE/T000341/1 (to RMDB, JPB, IG, DdV and MCJ). MK was supported by "la Caixa" Foundation (ID 100010434 to MK), fellowship code LCF/BQ/PR19/11700002 (to MK), and by the Vienna Science and Technology Fund (WWTF) and the City of Vienna through project VRG20-001 (to MK). JDO was supported by "la Caixa" Foundation (ID 100010434 to JDO) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 847648 (to JDO). The fellowship code is LCF/BQ/PI20/11760004 (to JDO). FES was supported by Brazilian National Council for Scientific and Technological Development (CNPq) (Process numbers.: 200502/2015-8, 302140/2020-4, 300365/2021-7, 301407/2021-5, 301925/2021-6 to FES), and received funding from International Primatological Society - Conservation grant; The Rufford Foundation (14861-1, 23117-2), the Margot Marsh Biodiversity Foundation (SMA-CCO-G0000000023, SMA-CCOG0000000037), Primate Conservation Inc. (#1713 and #1689), and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801505 (to FES). The Mamirauá Institute for Sustainable Development received funds from Gordon and Betty Moore Foundation (Grant #5344 to FES). Fieldwork for samples collected in the Brazilian Amazon was funded by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/SISBIOTA Program #563348/2010-0 to IPF), Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM/SISBIOTA #2317/2011 to IPF), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES AUX # 3261/2013) to IPF. Sampling of nonhuman primates in Tanzania was funded by the German Research Foundation (KN1097/3-1 to SK and RO3055/2-1 to CR) and by the US National Science Foundation (BNS83-03506 to JPC). No animals in Tanzania were sampled purposely for this study. Details of the original study on Treponema pallidum infection can be requested from SK. Sampling of baboons in Zambia was funded by US NSF grant BCS-1029451 to JPC, CJJ and JR. The research reported in this manuscript was also funded by the Vietnamese Ministry of Science and Technology's Program 562 (grant no. ĐTĐL.CN-64/19). ANC is supported by AEI-PGC2018-101927-BI00 704 (FEDER/UE to ANC), FEDER (Fondo Europeo de Desarrollo Regional)/FSE (Fondo Social Europeo), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M to ANC) and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880 to ANC). ADM was supported by the National Sciences and Engineering Research Council of Canada and Canada Research Chairs program. The authors would like to thank the Veterinary and Zoology staff at Wildlife Reserves Singapore for their help in obtaining the tissue samples, as well as the Lee Kong Chian Natural History Museum for storage and provision of the tissue samples. We wish to thank H. Doddapaneni, D.M. Muzny and M.C. Gingras for their support of sequencing at the Baylor College of Medicine Human Genome Sequencing Center. We greatly appreciate the support of Richard Gibbs, Director of HGSC for this project and thank Baylor College of Medicine for internal funding. TMB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant

agreement No. 864203 to TMB), BFU2017-86471-P (MINECO/FEDER, UE to TMB), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M to TMB), NIH 1R01HG010898-01A1 (to TMB) and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2021 SGR 00177 to TMB), Howard Hughes International Early Career (to TMB), Obra Social "La Caixa" and internal funds from Baylor College of Medicine. HLR receives funding from Illumina, Inc to support rare disease gene discovery and diagnosis. JPB, RMDB, IG and DV were supported by a UKRI Grant NERC (NE/T000341/1). We thank Dr. Praveen Karanth (IISc), Dr. H.N. Kumara (SACON) for collecting and providing us with some of the samples from India. SMA was supported by a BINC fellowship from the Department of Biotechnology (DBT), India. We acknowledge the support provided by the Council of Scientific and Industrial Research (CSIR), India to GU for the sequencing at the Centre for Cellular and Molecular Biology (CCMB), India. Aotus azarae samples from Argentina where obtained with grant support to EFD from the Zoological Society of San Diego, Wenner-Gren Foundation, the L.S.B. Leakey Foundation, the National Geographic Society, the U.S. National Science Foundation (NSF-BCS-0621020, 1232349, 1503753, 1848954; NSF-RAPID-1219368, NSF-FAIN-1952072; NSF-DDIG-1540255; NSF-REU 0837921, 0924352, 1026991) and the U.S. National Institute on Aging (NIA- P30 AG012836-19, NICHD R24 HD-044964-11). JHS was supported in part by the NIH under award number P40OD024628 - SPF Baboon Research Resource. This research is supported by the National Research Foundation Singapore under its National Precision Medicine Programme (NPM) Phase II Funding (MOH-000588 to PT and WKL) and administered by the Singapore Ministry of Health's National Medical ResearchCouncil. JR is also a Core Scientist at the Wisconsin National Primate Research Center, Univ. of Wisconsin, Madison. We acknowledge the institutional support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and the 2014–2020 Smart Growth Operating Program, to the EMBL partnership and institutional co-financing with the European Regional Development Fund (MINECO/FEDER, BIO2015-71792-P). We also acknowledge the support of the Centro de Excelencia Severo Ochoa, and the Generalitat de Catalunya through the Departament de Salut, Departament d'Empresa i Coneixement and the CERCA Programme to the institute.

Author contributions: HG, TH, JE, JGS, JM, MSB, YY, AD, PF, LK, LS, YW, AA, YF, SC, SB, GL, RR, DB, FA, KF performed the analysis and wrote the manuscript. MCJ, MK, JDO, SM, AV, JB, MR, FES, LA, JB, MG, DdV, IG, RAH, MR, AJ, ISC, JH, CH, DJ, PF, FRdM, FB, HB, IS, IF, JVdA, MM, MNFdS, MT, RR, TH, NA, CJR, AZ, CJJ, JPC, GW, CA, JHS, EFD, SK, FS, DW, LZ, YS, GZ, JDK, SK, MDL, EL, SM, AN, TB, TN, CCK, JL, PT, WKL, ACK, DZ, IG, AM, KG, MHS, RMDB, GU, CR, JPB contributed the primate samples and sequencing data. ML, SS, AOD, HR, JX, JR, TMB, and KF supervised the work.

1. **Competing interests:** Employees of Illumina, Inc. are indicated in the list of author affiliations. Serafim Batzoglou is currently affiliated with Seer, Inc. Heidi Rehm receives funding to support rare disease research and tool development from Illumina, Inc. and Microsoft, Inc. Patents related to this work are (1) title: Deep convolutional neural networks to predict variant pathogenicity using three-dimensional (3D) protein structures, filing No.: US 17/232,056, authors: Tobias Hamp, Kai-How Farh, Hong Gao; (2) title:

Transfer learning-based use of protein contact maps for variant pathogenicity prediction, filing No.: US 17/876,481, authors: Chen Chen, Hong Gao, Laksshman Sundaram, Kai-How Farh; (3) title: Multi-channel protein voxelization to predict variant pathogenicity using deep convolutional neural networks, filing No.: US 17/703,935, authors: Tobias Hamp, Kai-How Farh, Hong Gao;(4) title: Transformer language model for variant pathogenicity, filing No.: US 17/975,536 and US 17/975,547, authors: Jeffrey Ede, Tobias Hamp, Anastasia Dietrich, Yibing Wu, Kai-How Farh.

Data and materials availability: All sequencing data have been deposited at the European Nucleotide Archive under the accession number PRJEB49549. Primate variants and PrimateAI-3D prediction scores are available with a non-commercial license upon request and are displayed on https://primad.basespace.illumina.com. The source code of PrimateAI-3D is accessible via https://github.com/Illumina/PrimateAI-3D and is also archived at https://doi.org/10.5281/zenodo.7738731. To reduce problems with circularity that have become a concern for the field, the authors explicitly request that the prediction scores from the method not be incorporated as a component of other classifiers, and instead ask that interested parties employ the provided source code and data to directly train and improve upon their own deep learning models.

Supplementary Materials:

Materials and Methods

Supplementary Text

Figs. S1 to S28

Tables S1 to S6

References (*125-169*)

Figures

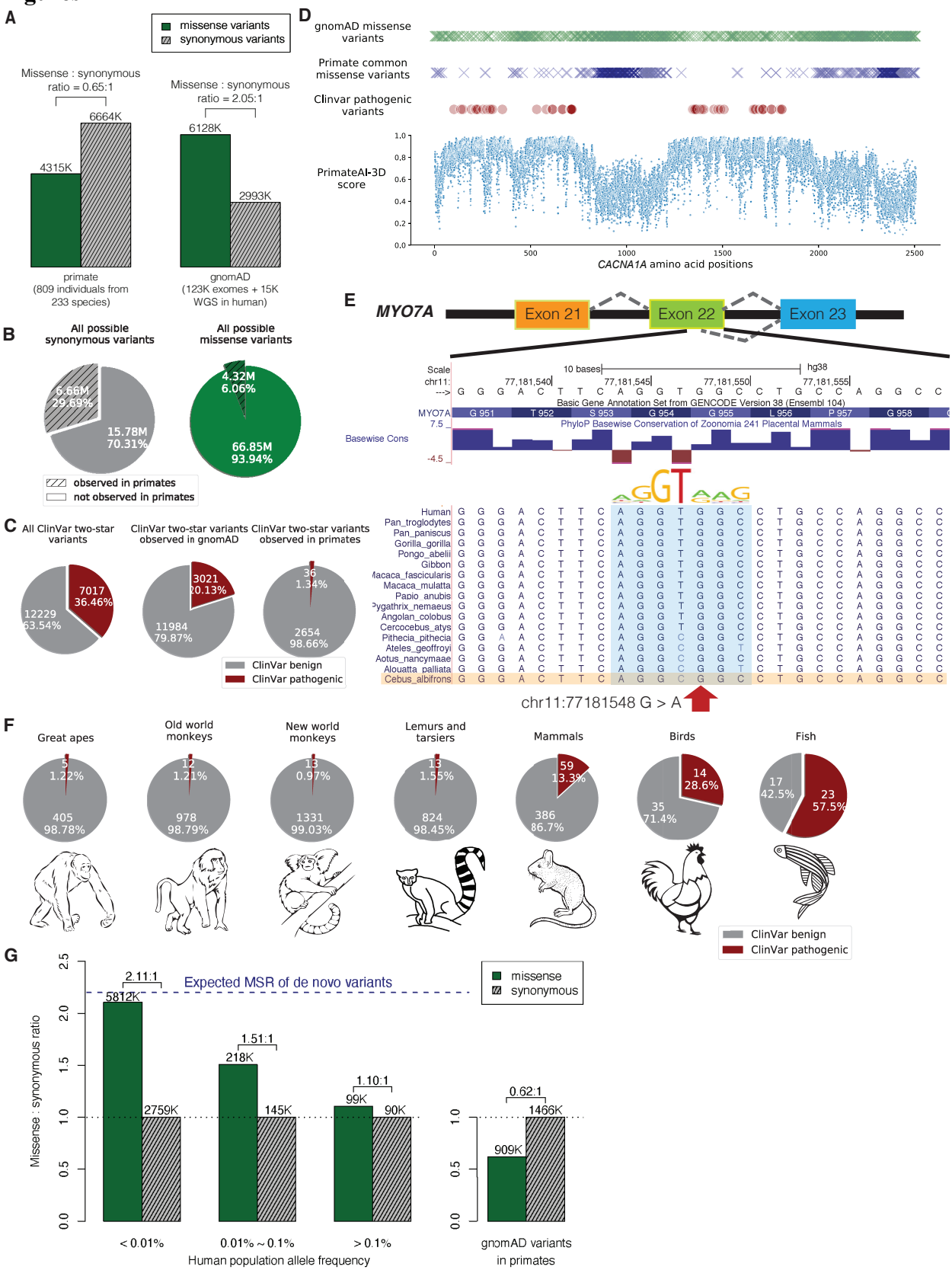
A
missense variants
synonymous variants
Missense : synonymous ratio = 0.65:1
6664K
4315K
Missense : synonymous ratio = 2.05:1
6128K
2993K
primate (809 individuals from 233 species)
gnomAD (123K exomes + 15K WGS in human)
B
All possible synonymous variants
All possible missense variants
5.66M 29.69%
15.78M 70.31%
4.32M 6.06%
66.85M 93.94%
observed in primates
not observed in primates
C
All ClinVar two-star variants
ClinVar two-star variants observed in gnomAD
ClinVar two-star variants observed in primates
7017 36.46%
12229 63.54%
3021 20.13%
11984 79.87%
36 1.34%
2654 98.66%
ClinVar benign
ClinVar pathogenic
D
gnomAD missense variants
Primate common missense variants
Clinvar pathogenic variants
PrimateAI-3D score
CACNA1A amino acid positions
E
MYO7A
Exon 21
Exon 22
Exon 23
Basic Gene Annotation Set from GENCODE Version 38 (Ensembl 104)
PhyloP Basewise Conservation of Zoonomia 241 Placental Mammals
Basewise Cons
chr11:77181548 G > A
F
Great apes
Old world monkeys
New world monkeys
Lemurs and tarsiers
Mammals
Birds
Fish
5 1.22%
405 98.78%
12 1.21%
978 98.79%
13 0.97%
1331 99.03%
13 1.55%
824 98.45%
59 13.3%
386 86.7%
14 28.6%
35 71.4%
17 42.5%
23 57.5%
ClinVar benign
ClinVar pathogenic
G
Missense : synonymous ratio
Expected MSR of de novo variants
missense
synonymous
2.11:1
5812K
2759K
1.51:1
218K
145K
1.10:1
99K
90K
0.62:1
909K
1466K
< 0.01%
0.01% ~ 0.1%
> 0.1%
Human population allele frequency
gnomAD variants in primates

Fig. 1. Common primate variants are largely benign in human. (**A**) Counts of missense (solid green) and synonymous (shaded grey) variants from primates compared to the gnomAD database. Missense : synonymous counts and ratios are displayed above each bar. (**B**) Fractions of all possible human synonymous (grey) and missense variants (green) observed in primates. (**C**) Counts of benign (grey) and pathogenic (red) missense variants with two-star review status or above in the overall ClinVar database (left pie chart), compared to ClinVar variants observed in gnomAD (middle), and compared to ClinVar variants observed in primates (right). Conflicting benign and pathogenic annotations and variants interpreted only with uncertain significance were excluded. (**D**) Observed gnomAD (green) or primate (blue) missense variants in each amino acid position in the *CACNA1A* gene. Red circles represent the positions of annotated ClinVar pathogenic missense variants. Bottom scatterplot shows PrimateAI-3D predicted pathogenicity scores for all possible missense substitutions along the gene. (**E**) Multiple sequence alignment showing the ClinVar pathogenic variant chr11:77181548 G>A (red arrow) creating a cryptic splice site in human sequence (extended splice motif, blue). This variant is tolerated in Cebus Albifrons and other species with a G>C synonymous change in the adjacent nucleotide that stops the splice motif from forming. (**F**) Pie charts showing the fraction of benign (grey) and pathogenic (red) missense variants with ClinVar two-star review status or above in great apes, old world monkeys, new world monkeys, lemurs/tarsiers, mammals, chicken, and zebrafish. (**G**) Missense : synonymous ratios across the human allele frequency spectrum, with MSR of human variants seen in primates shown for comparison. The blue dashed line represents the expected missense : synonymous ratio of de novo variants. Colors and legend are the same as (**A**).

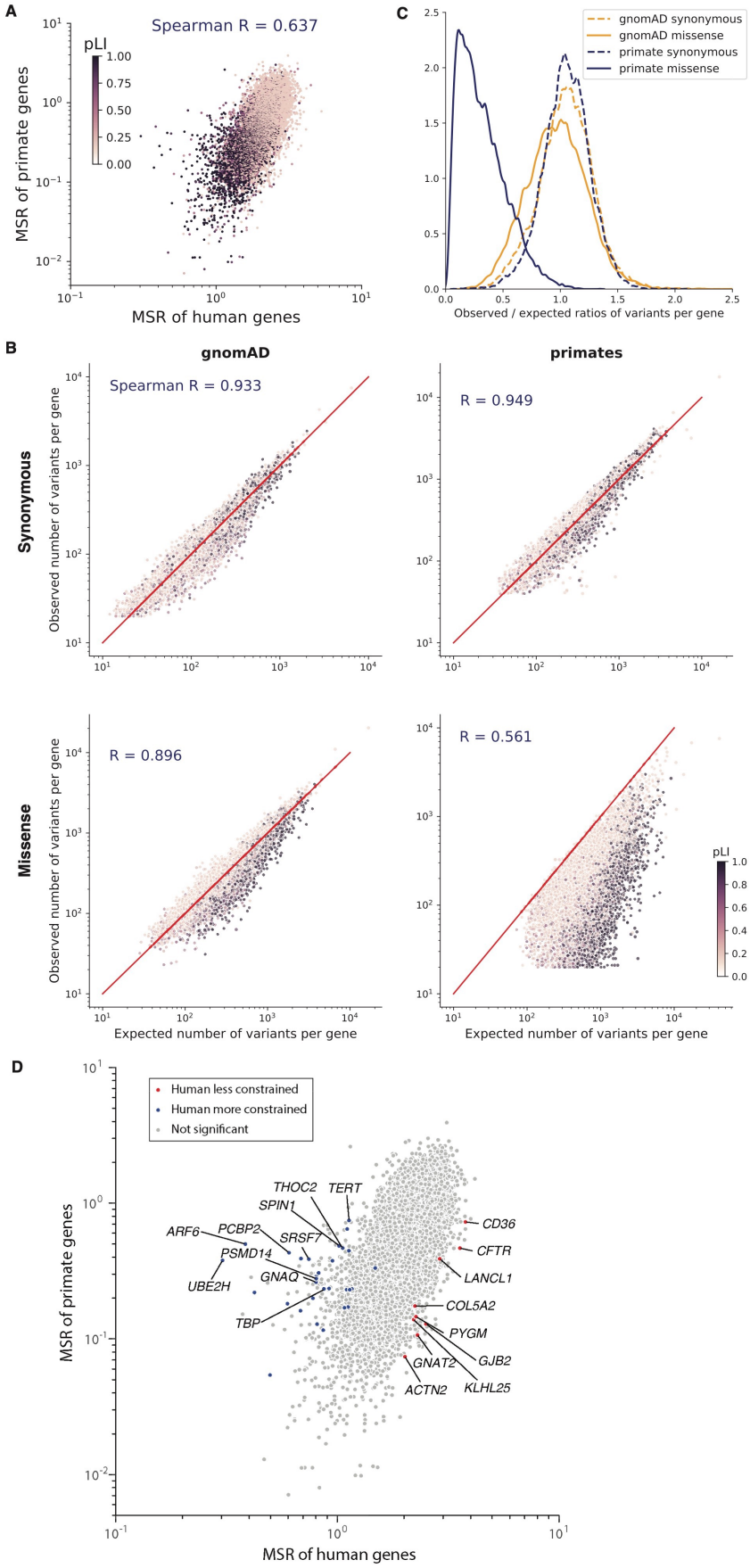
A
Spearman R = 0.637
pLI
1.00
0.75
0.50
0.25
0.00
MSR of primate genes
MSR of human genes
C
gnomAD synonymous
gnomAD missense
primate synonymous
primate missense
Observed / expected ratios of variants per gene
B
gnomAD
primates
Synonymous
Missense
Spearman R = 0.933
R = 0.949
R = 0.896
R = 0.561
Observed number of variants per gene
Expected number of variants per gene
pLI
D
Human less constrained
Human more constrained
Not significant
THOC2
TERT
SPIN1
ARF6
PCBP2
SRSF7
PSMD14
GNAQ
UBE2H
TBP
CD36
CFTR
LANCL1
COL5A2
PYGM
GJB2
GNAT2
ACTN2
KLHL25
MSR of primate genes
MSR of human genes

Fig. 2. Selective constraint of primate genes compared to human. (**A**) Scatter plot of missense : synonymous ratios between primate and human genes. Each gene is colored by its pLI score, with darker points showing haploinsufficient genes. (**B**) Observed and expected counts of synonymous (top) and missense (bottom) variants per gene in gnomAD (left) and primates (right). Genes are colored by their pLI scores. (**C**) Distributions of observed/expected ratios of synonymous (dashed lines) and missense (solid lines) variants for all genes. Results for primate genes (orange) and gnomAD genes (blue) are shown. (**D**) Scatter plot of missense : synonymous ratios between primate and human genes. Highlighted points are genes that are under significantly stronger (blue) or weaker (red) constraint in humans compared to non-human primates under both methods (Benjamini-Hochberg FDR < 0.05), while grey points show non-significant genes. The top 10 genes with the largest effect sizes in either direction are labeled.

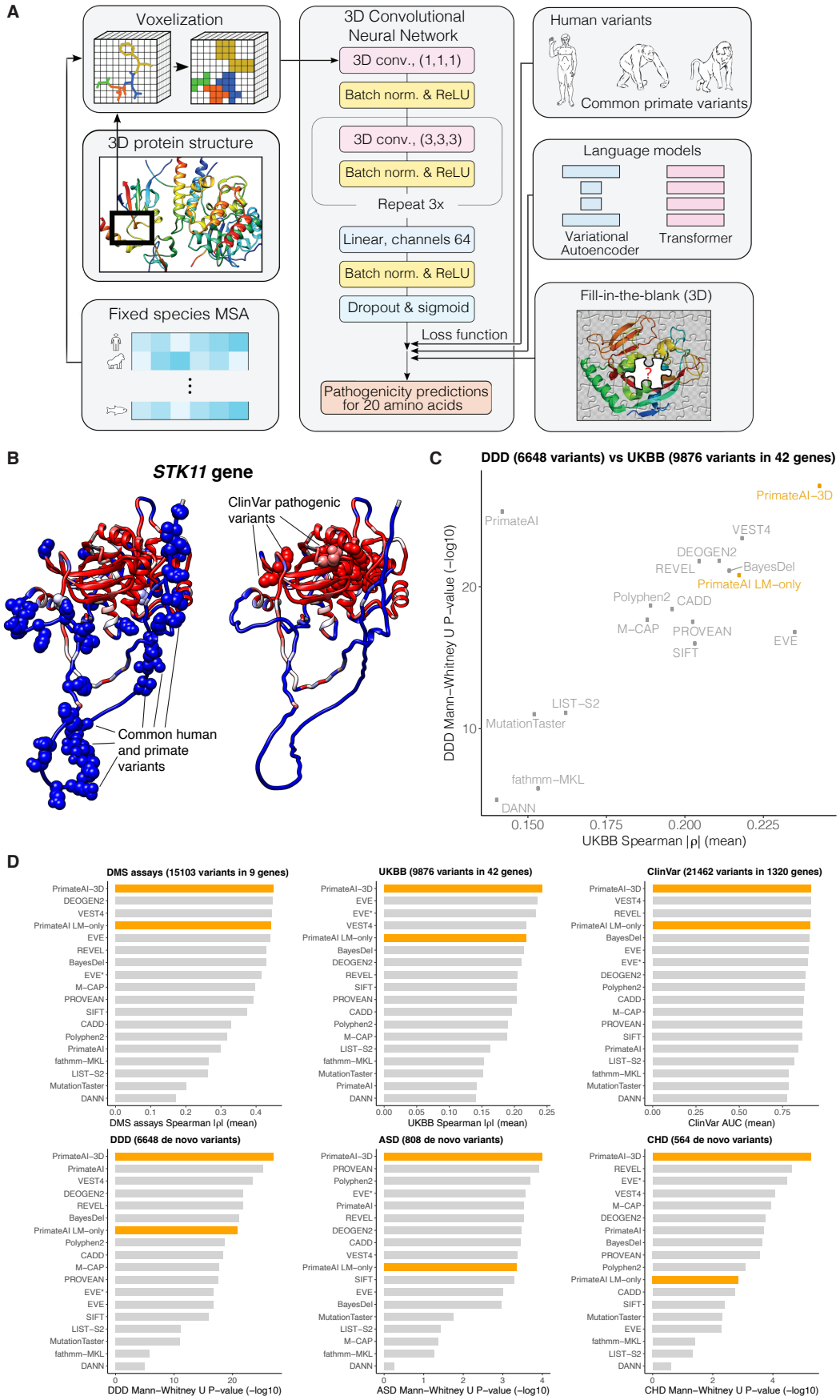
A
Voxelization
3D protein structure
Fixed species MSA
3D Convolutional Neural Network
3D conv., (1,1,1)
Batch norm. & ReLU
3D conv., (3,3,3)
Batch norm. & ReLU
Repeat 3x
Linear, channels 64
Batch norm. & ReLU
Dropout & sigmoid
Loss function
Pathogenicity predictions for 20 amino acids
Human variants
Common primate variants
Language models
Variational Autoencoder
Transformer
Fill-in-the-blank (3D)
B
STK11 gene
ClinVar pathogenic variants
Common human and primate variants
C
DDD (6648 variants) vs UKBB (9876 variants in 42 genes)
DDD Mann−Whitney U P−value (−log10)
UKBB Spearman |ρ| (mean)
PrimateAI−3D
PrimateAI
VEST4
DEOGEN2
REVEL
BayesDel
PrimateAI LM−only
Polyphen2
CADD
M−CAP
PROVEAN
SIFT
EVE
LIST−S2
MutationTaster
fathmm−MKL
DANN
D
DMS assays (15103 variants in 9 genes)
DMS assays Spearman |ρ| (mean)
UKBB (9876 variants in 42 genes)
UKBB Spearman |ρ| (mean)
ClinVar (21462 variants in 1320 genes)
ClinVar AUC (mean)
DDD (6648 de novo variants)
DDD Mann−Whitney U P−value (−log10)
ASD (808 de novo variants)
ASD Mann−Whitney U P−value (−log10)
CHD (564 de novo variants)
CHD Mann−Whitney U P−value (−log10)

Fig. 3. PrimateAI-3D architecture and variant classification performance. (**A**) PrimateAI-3D workflow. Human protein structures and multiple sequence alignments are voxelized (left) as input to a 3D convolutional neural network that predicts pathogenicity of all possible point mutations of a target residue (middle). The network is trained using a loss function with three components (right): common human and primate variants; fill-in-the-blank of a protein structure; score ranks from language models. (**B**) Protein structure of the *STK11* gene, colored by PrimateAI-3D pathogenicity prediction scores (blue: benign; red: pathogenic). Spheres indicate residues with common human and primate variants (left) or residues with pathogenic mutations from ClinVar (right). For spheres, the color corresponds to the pathogenicity score of only the variant. For other residues, pathogenicity scores are averaged over all variants at that site. (**C**) Scatterplot shows performance of methods that predict missense variant pathogenicity in two clinical benchmarks (DDD and UKBB). Datasets are a subset of variants for which all methods have predictions. (D) Six barplots show method performance for six testing datasets (DMS assays, UKBB, ClinVar, DDD, ASD, and CHD).

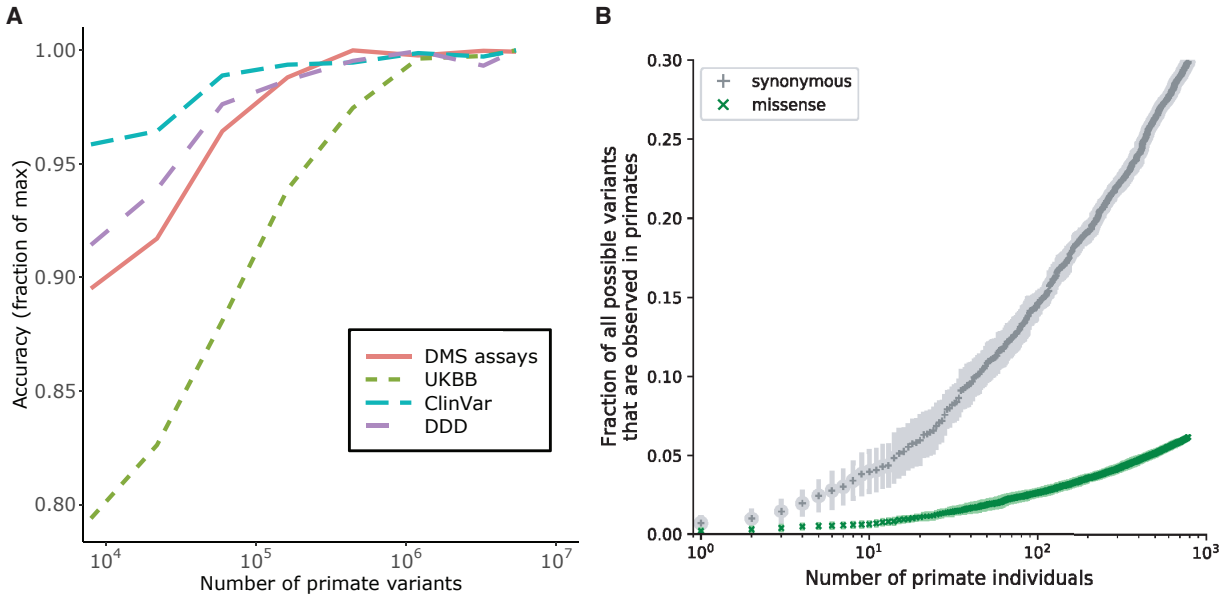

Fig. 4. Impact of training dataset size on classification accuracy. (**A**) Improved performance of PrimateAI-3D with increasing number of common human and primate variants in the training dataset (x-axis). Performance of each dataset (y-axis) was divided by the maximum performance observed across all training dataset sizes. (**B**) Cumulative fractions of all possible human synonymous (grey) and missense (green) variants observed as common variants in 234 primate species, including human (allele frequency > 0.1%). Each point shows the average of ten permutations, calculated with a different random ordering of the list of primate species each time.

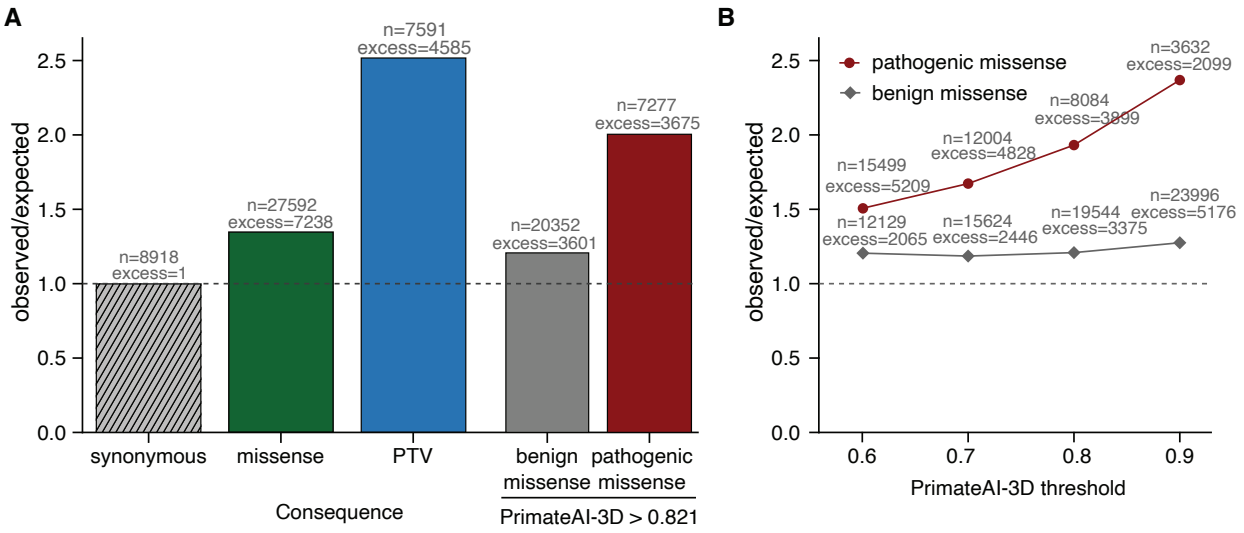

Fig. 5. Enrichment of de novo mutations in the neurodevelopmental disorder cohort over expectation. (**A**) Enrichment of DNMs from Kaplanis *et al. (87)* across all genes. Enrichment ratios are given for synonymous, all missense, and protein-truncating variants (PTV), along with missense split by PrimateAI-3D score into benign (<0.821) and pathogenic (>0.821). (**B**) Enrichment of benign and pathogenic missense above expectation at varying PrimateAI-3D thresholds for classifying pathogenic missense.

HGNC symbol	Protein-truncating variants	Missense		*P* value	
		PrimateAI-3D score ≥ 0.821	All missense	PrimateAI-3D score ≥ 0.821	All missense
AP1G1	2	4	5	4.1×10^{-7}	5.9×10^{-5}
ATP2B2	1	9	11	2.1×10-^{7}	1.4×10^{-3}
CELF2	2	4	4	1.2×10^{-7}	6.7×10^{-5}
MAP4K4	2	6	7	3.9×10^{-7}	5.0×10^{-4}
MED13	3	6	9	6.6×10^{-8}	3.5×10^{-5}
MFN2	0	6	8	3.4×10^{-7}	1.0×10^{-5}
NR4A2	2	4	5	3.7×10^{-7}	3.3×10^{-5}
PIP5K1C	0	8	9	2.8×10^{-8}	4.9×10^{-4}
RAB5C	2	4	5	8.6×10^{-8}	1.5×10^{-5}
SPOP	1	4	6	4.1×10^{-7}	1.7×10^{-6}
SPTBN2	1	10	16	3.9×10^{-7}	4.5×10^{-3}
XPO1	1	7	7	5.0×10^{-7}	7.2×10^{-4}
EIF4A2	2	4	4	1.7×10^{-7}	2.1×10^{-4}
LMBRD2	0	3	4	6.0×10^{-7}	1.3×10^{-4}
MARK2	4	3	5	2.3×10^{-7}	3.8×10^{-5}
NOTCH1	4	6	17	4.1×10^{-7}	1.3×10^{-6}

Table 1. Additional genes discovered in intellectual disability. Genes achieving the genome-wide significance ($p < 6.4\text{x}10^{-7}$) are shown when considering only missense de novo mutations with PrimateAI-3D scores ≥0.821. Counts of protein truncating and missense DNMs are provided. *P* values for gene enrichment are shown when the statistical test was run only with missense mutations with PrimateAI-3D score ≥ 0.821, and when it was repeated for all missense mutations.