nature biotechnology

ARTICLES

Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases

Rong Chen[1,2,12], Lisong Shi[1,2,12], Jörg Hakenberg[1,2], Brian Naughton[3,11], Pamela Sklar[1,2,4], Jianguo Zhang[5], Hanlin Zhou[5], Lifeng Tian[6], Om Prakash[7], Mathieu Lemire[8], Patrick Sleiman[6], Wei-yi Cheng[1,2], Wanting Chen[5], Hardik Shah[1,2], Yulan Shen[5], Menachem Fromer[1,2,4], Larsson Omberg[9], Matthew A Deardorff[6], Elaine Zackai[6], Jason R Bobe[1,2], Elissa Levin[1,2], Thomas J Hudson[8], Leif Groop[7], Jun Wang[10], Hakon Hakonarson[6], Anne Wojcicki[3], George A Diaz[1,2], Lisa Edelmann[1,2], Eric E Schadt[1,2] & Stephen H Friend[1,2,9]

Genetic studies of human disease have traditionally focused on the detection of disease-causing mutations in afflicted individuals. Here we describe a complementary approach that seeks to identify healthy individuals resilient to highly penetrant forms of genetic childhood disorders. A comprehensive screen of 874 genes in 589,306 genomes led to the identification of 13 adults harboring mutations for 8 severe Mendelian conditions, with no reported clinical manifestation of the indicated disease. Our findings demonstrate the promise of broadening genetic studies to systematically search for well individuals who are buffering the effects of rare, highly penetrant, deleterious mutations. They also indicate that incomplete penetrance for Mendelian diseases is likely more common than previously believed. The identification of resilient individuals may provide a first step toward uncovering protective genetic variants that could help elucidate the mechanisms of Mendelian diseases and new therapeutic strategies.

Advances in genomic technologies have rapidly expanded our knowledge of the genetic basis of human disease. To date, >6,000 Mendelian disorders have been described (Online Mendelian Inheritance in Man (OMIM)[1]), with more than 150,000 disease-associated variants identified across these disorders in the Human Gene Mutation Database (HGMD)[2]. Despite the success of genome-wide association and whole-exome and whole-genome sequencing (WES/WGS) studies in revealing the DNA variants that underlie the genetic basis of disease, the development of effective treatments for most diseases has remained a challenge. Even for Mendelian disorders, only a handful of drugs have been developed[3]. One reason for this lack of success is the difficulty of using small-molecule therapies to restore protein activity in the presence of loss-of-function (LoF) mutations. As a result, treatment of Mendelian disorders typically focuses on the relief of symptoms rather than on a biological 'cure'.

A promising avenue for addressing some of these limitations is to focus analysis on the genetic and environmental modulators that keep people well by suppressing the effects of disease-causing mutations[4]. However, a major challenge in identifying resilient individuals is accurately cataloging disease mutations. Currently, there are no databases that provide a complete characterization of disease genes and their mutations as well as in-depth clinical annotations. For example, the OMIM[1] database contains all known Mendelian disorders with detailed clinical characterizations, but has limited descriptions of disease-causing mutations. In contrast, HGMD[2] has collected almost all disease-associated variants reported to date, but has almost no parameters pertaining to the clinical characteristics attributed to these variants. Furthermore, although many commercial pan-ethnic screening panels cover the most common highly penetrant mutations[5–7], important mutations might be omitted owing to technological limitations and cost-benefit considerations. Also, the exact mutations in these commercial pan-ethnic screening panels are typically inaccessible to the public.

Despite these challenges, identification of secondary modulators has proven successful across a multitude of model organisms in which the prominent role of second-site suppressors that buffer or modify traits has been established[8–11]. For example, human genetic studies have identified rare mutations in *CCR5* that confer resilience against HIV infection[12], mutations in globin genes that modify the severity of sickle cell disease by buffering primary mutations in β-globin genes[13], and LoF mutations in *PCSK9* that protect carriers from high lipid levels and resulting heart disease[14]. Second-site mutations in disease genes have also been shown to revert clinical phenotype in patients with recessive dystrophic epidermolysis[15] and Fanconi anemia[16], whereas LoF mutations in zinc transporter 8 have been found to protect obese individuals from diabetes[17]. Most recently, a variant

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [3]23andMe, Mountain View, California, USA. [4]Friedman Brain Institute and Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [5]BGI-Shenzhen, Shenzhen, China. [6]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [7]Department of Clinical Sciences, Diabetes & Endocrinology, Lund University Diabetes Center, Skåne University Hospital, Lund University, Malmö, Sweden. [8]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [9]Sage Bionetworks, Seattle, Washington, USA. [10]iCarbonX, Shenzhen, China. [11]Present address: Boolean Biotech Inc., Mountain View, California, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to E.E.S. (eric.schadt@mssm.edu) or S.H.F. (friend@sagebase.org) or R.C. (rong.chen@mssm.edu).

Received 29 July 2015; accepted 12 February 2016; published online 11 April 2016; corrected online 21 April 2016; doi:10.1038/nbt.3514

identified in the gene *Jagged1* was found to confer resilience to Duchenne muscular dystrophy in two dogs, implicating *Jagged1* as a therapeutic target for the disorder[18].

Here we analyze sequence and genotype data from 589,306 individuals across 12 studies (complete list in Online Methods) to identify healthy individuals harboring what are currently believed to be completely penetrant Mendelian disease-causing mutations. We refer to this search for resilient individuals as the Resilience Project. We screen mutations in 874 genes believed to cause 584 distinct severe Mendelian childhood disorders. In total, we identified 13 candidate resilient individuals spanning 8 diseases. The genomes of such resilient individuals, if appropriately decoded, hold promise in elucidating protective mechanisms of disease that could lead to novel treatments[19].

RESULTS

We carried out a search of existing genomic data for individuals who may be resilient to disease by focusing on mutations annotated as being completely penetrant for severe childhood Mendelian disorders. Our rationale for restricting attention to these disorders is manifold. First, there is a significant unmet medical need for many of these disorders that have the potential to benefit from the identification of resilient individuals. Second, a focus on diseases with a more profound phenotype and a simple genetic architecture decreased the chances of diagnostic errors or missed diagnoses due to subclinical manifestation of disease. This is particularly important for our screen, given we generally did not have access to medical records and depended on self-reporting of conditions by study participants. Finally, restricting attention to severe childhood disorders and including only individuals over the age of 18 reduces the likelihood that subjects harboring deleterious mutations will manifest the disorder later in life. The overall workflow for the retrospective search for resilient individuals is depicted in **Figure 1**.

Building gene and allele panels

The search for individuals who are resilient to severe childhood disorders required the construction of a screening panel of alleles known to cause such disorders with complete penetrance (**Supplementary Fig. 1**). A multi-stage filter was applied to identify the subset of disorders that fit our criteria. Diseases annotated as mild or of unknown severity, with an unknown age of onset or an age of onset later than 18 years, or with incomplete or unknown penetrance were removed, leaving 584 unique Mendelian diseases spanning 17 different disease categories and 874 implicated genes. This comprised the disease gene panel for our study (**Table 1** and **Supplementary Table 1**). The top three most-represented disease categories were metabolic conditions, neurological diseases and developmental disorders, which accounted for 22.9%, 16.8% and 15.6% of the disease genes, respectively.

Disease-causing mutations in genes in the disease gene panel were identified using two independent pipelines. The first, comprising a core allele panel (CAP; **Supplementary Table 2**), aimed to identify well-established and well-annotated disease mutations, and the second, comprising an expanded allele panel (EAP), aimed to identify mutations that have strong support for causing severe childhood disorders. The CAP comprised 674 founder or major recurrent mutations from 162 genes representing 125 severe, early-onset diseases. Among these mutations, 47% were missense, 20% were nonsense, 11% affected splicing, 4% were in-frame insertions or deletions, and

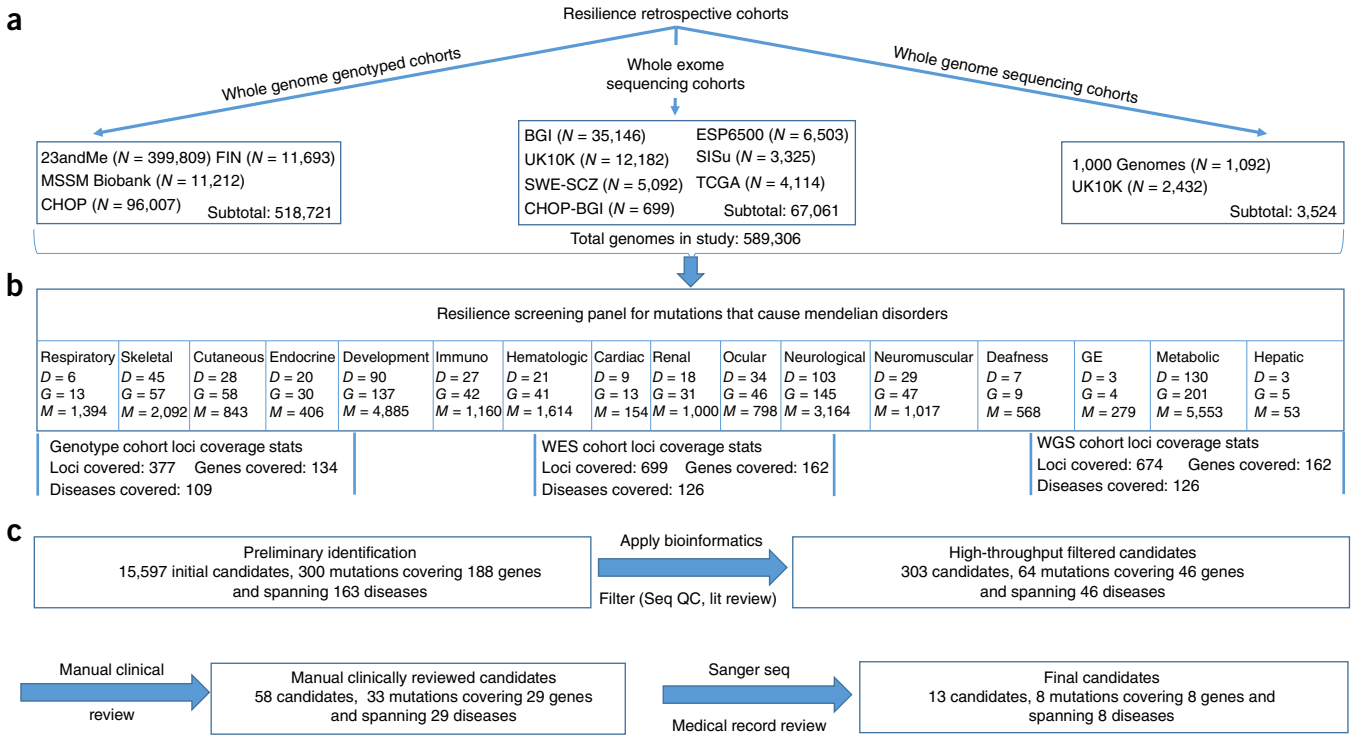

Figure 1 Study design and results for the retrospective search for resilient individuals. (**a**) A summary of the different cohorts and the genomic data available on those cohorts (see **Table 2** for more details). (**b**) The disease-causing genes and mutations that were assembled to construct our screening panel (more details in **Table 1** and **Supplementary Tables 1** and **2**). The D, G and M variables denote the number of diseases, genes and mutations, respectively, represented on our screening panel in the respective disease categories. The coverage statistics indicate the coverage achieved for the core allele panel in the genotype, WES and WGS cohorts. (**c**) Summaries for the different stages of the filtering process to identify candidate resilient individuals (see **Supplementary Fig. 1** and **Tables 3** and **4** for more details).

Table 1 The Resilience Project gene and allele panels cover diseases from 16 categories

Disease category	Core allele panel			Gene panel	
	No. of diseases	No. of genes	No. of mutations	No. of diseases	No. of genes
Cardiac	0	0	0	9	13
Cutaneous	3	6	24	28	58
Deafness	0	0	0	7	9
Developmental	11	19	53	90	137
Endocrine	1	1	6	20	30
Gastroenterological	0	0	0	3	4
Hematologic	2	2	6	24	41
Hepatic	0	0	0	3	5
Immunodeficiency	6	9	19	27	42
Metabolic	64	78	316	130	201
Neuromuscular	5	8	24	29	47
Neurological	21	27	74	105	147
Ocular	3	3	15	34	46
Renal	4	4	23	18	31
Respiratory	1	1	110	6	13
Skeletal	6	7	15	45	57
Other	0	0	0	8	8
Total	125	162[a]	674	584	874[a]

[a]Does not equal the total number in the column since some genes are associated with multiple diseases from different categories.

the remaining 18% were frameshift insertions or deletions resulting in premature stop codons (**Supplementary Fig. 2**). The EAP was intended to complement the CAP by casting a broader net for disease mutations in genes in the disease gene panel, tolerating a higher number of false positives with respect to our selection criteria for the initial identification of resilient individuals, and resolving the false-positive identifications by manual curation and clinical review. The EAP covered 24,186 variants from HGMD tagged as "disease causing mutations" (DM) with allele frequencies lower than 0.5% in the 1000 Genomes Project[20] and NHLBI GO Exome Sequencing Project (ESP)6500 (ref. 21; **Table 1**).

Applying CAP and EAP to screen 589,306 genomes

In our search for resilient individuals, we analyzed existing DNA sequence and genotype data from 12 past and ongoing genetic studies worldwide (Online Methods and **Table 2**). Combined, these data sets provided genome-wide variant data on 589,306 individuals. Because individual-level data could not be shared across studies, we were unable to definitively assess the number of unique individuals represented. However, we anticipate that all 589,306 individuals are unique given the geographic separation between most of the studies and the low sampling rates in the studies that sampled across broader geographic regions. We verified this in the samples from 2 of the 12 studies, 1000 Genomes and UK10K project[22] samples using a single-nucleotide polymorphism (SNP) panel of 40 polymorphic markers. In comparing all samples pairwise across these two studies, we identified no duplicate samples, in addition to 18 twin pairs from UK10K.

Given the different genotyping or sequencing assays run across the cohorts in our study, the coverage across all variants represented in CAP and EAP varied widely among the samples (**Supplementary Fig. 3**). A subset of 59 loci in CAP was covered across all samples in the study. For The Cancer Genome Atlas (TCGA) Project, UK10K and 1000 Genomes studies, which comprised 19,820 samples, the assays covered all 674 loci in the CAP. However, for these data sets we did not obtain the per-sample coverage for each locus, so individual samples may not cover all loci. Per-sample coverage was available for only one cohort, the Swedish schizophrenia cohort (SWE-SCZ)[23]. These data were used to assess the extent of coverage achieved across all CAP loci. For the 5,092 samples in SWE-SCZ, 670 of the 674 loci in CAP are well-covered by all samples, with the remaining four loci having no coverage in any sample. The four loci not covered are intronic and are at least 20 nucleotides from the closest exon. For cohorts with genotype data, we used both assayed and imputed genotypes in the screen, making use of information on the quality of the called genotype, genotype likelihood and imputed genotype confidence to filter out spurious candidates. Of the 674 loci in CAP, the 23andMe, Mount Sinai BioBank, the Children's Hospital of Philadelphia (CHOP) BioBank and Finnish (components listed in Online Methods) cohorts had 297, 105, 59 and 163 filtered loci, respectively (**Supplementary Fig. 4**). Over all studies, the effective number of loci (as a proportion of all loci covered in CAP) was 36.5%.

Identifying candidate resilient individuals

We identified 15,597 candidate resilient individuals from our screen of 589,306 genomes against the CAP and EAP panels, representing 300 compound heterozygous or homozygous mutations across 188 genes for 163 Mendelian diseases. Of these 15,597 candidates, 367 were identified from the CAP (44 mutations), whereas the remaining 15,230 were identified from the EAP (256 mutations). We manually

Table 2 Data sources used in current retrospective study

Sample source	Sample type	Sample size	Technology	Population
TCGA	Matched normal tissues for 17 tumor types	4,114	WES and WGS	No population-specific data acquired
Mount Sinai BioBank	Various diseases	11,212	Genotyping array	Self-reported ethnicities
23andMe	Mixed	399,809	Genotyping array	No population-specific data acquired
1000 Genomes Projects	Healthy	1,092	Low pass WGS	African, American, Asian and European; subcategories available
ESP6500	Various diseases	6,503	WES	African-American and European-American (both USA)
UK10K[a]	Cohorts; neurodevelopmental disorders; obesity samples; rare diseases	14,614	Partly WGS, partly WES	Mostly UK and Finland; no population-specific data acquired
SISu[a,b]	Case-control mixed	3,325	WES	Finnish
FINN[a,c]	Case-control mixed	11,693	Genotyping array	Finnish
CHOP-BGI	Case-control mixed	699	WES	Mixed
CHOP	Case-control mixed	96,007	Genotyping array	Mixed
BGI	Case-control mixed	35,146	Partly WGS, partly WES	Mixed
SWE-SCZ	Schizophrenia cases and controls	5,092	WES	Swedish (some samples with partial Finnish ancestry)
Total WES/WGS		70,585		
Total genotyping		518,721		
Grand total		589,306		

[a]For detailed data, see **Supplementary Table 4**. [b]SISu, Sequencing Initiative Suomi (http://www.sisuproject.fi/): consortia including FINRISK, GoT2D (only the Fusion and Botnia studies), H2000, METSIM, NFBC66 and Finnish samples from the 1000 Genomes projects. [c]FINN, a subset of cohorts from SISu: FINRISK, EUFAM, Finnish Twin study and Migraine Study, with genome-wide genotype data.

reviewed all mutations represented in this group to ensure that the corresponding phenotype associated with these mutations met our criteria for inclusion (completely penetrant, severe phenotype, early age of onset) and to ensure the genotype calls were made with high confidence. We excluded 6,667 of 15,597 candidates due to low confidence in the genotype call as represented by either low sequencing depth, high GC or AT content, repetitive sequence region or skewed Hardy-Weinberg equilibrium statistics. We excluded an additional 8,627 candidates owing to high population frequency (>0.5%) of discovered variants or an inability to access individual data for follow-up (e.g., ESP data set) (**Table 3**).

For the remaining 303 candidates, we carried out a manual review of each mutation with a review team composed of bioinformatics scientists, board-certified clinical geneticists, medical consultants and genetic counselors to assess whether variation in the ages of onset and/or variations in the expression of the corresponding phenotype could explain why a candidate was flagged. For 245 of the 303 candidates, we determined the expressivity of the disease phenotype was not extreme enough to unambiguously categorize the candidate as completely resilient (**Table 3**). Another 16 candidates were excluded because the published literature could not provide sufficient evidence to support pathogenicity for the variants discovered in these individuals, although the diseases associated with the corresponding genes are generally severe enough to be considered as candidates in our list.

After reviewing available medical records for the remaining 42 candidates, 14 presented expected manifestations from the genotypes they carried, indicating that they did not meet the criteria of a 'healthy' individual. Sanger sequencing ruled out another 15 candidates because the genotypes were determined to be heterozygous, not homozygous, as originally determined from the variant data. The final 13 candidates all harbored homozygous (autosomal recessive disease) or heterozygous (autosomal dominant disease) mutations to one of eight different severe Mendelian childhood disorders that would normally be expected to cause severe disease before the age of 18 years: cystic fibrosis, Smith-Lemli-Opitz syndrome, familial dysautonomia, epidermolysis bullosa simplex, Pfeiffer syndrome, autoimmune polyendocrinopathy syndrome, acampomelic campomelic dysplasia and atelosteogenesis (**Table 4**; **Table 5** and **Supplementary Fig. 5**). The severity of the expected phenotypes makes it highly unlikely that such an individual would have manifested the disease without it being clearly annotated in their health records. A review of the individual health information for six candidates was performed, and no evidence of the indicated disease was uncovered. Genotypes for 5 of the 13 candidates were confirmed by Sanger sequencing to be true homozygotes, whereas the remaining 8 candidates from the UK10K[22], 23andMe, Sequencing Initiative Suomi or SISu (http://www.sisuproject.fi/), and BGI cohorts could not be validated owing to insufficient remaining DNA for these samples.

We modeled estimates regarding the number of expected resilient individuals from our study cohort with all autosomal recessive alleles in CAP, based on allele frequencies in the ExAC[24], DIVAS[25] and related databases and penetrance information (**Supplementary Table 3**). We estimated that we would have expected to identify 9 or 10 individuals with the indicated genotype out of all of those screened, which is not significantly different from the number of candidates we identified ($P > 0.05$).

Table 3 Reasons for filtering out initial candidates due to sequencing quality, inaccurate information obtained from databases, clinical review of mutations, and clinical review of individual medical record

Reason	Secondary reason	Annotation	No. of mutations	No. of diseases	No. of individuals	Example	Reference/data source
Sequencing quality	Low coverage	Average coverage <10	59	38	3,383	ZNF469 - c.1541_1542insG	EVS
	High GC or AT	5′ or 3′ UTR	5	5	7	PEX1 - c.523_524insG	GRCh37/hg19
	Repetitive sequence	Homopolymer, tandem repeats, genomic segmental duplication	9	8	15	PYGM - c.2262delA	GRCh37/hg19
	Genotype calling mistake	Miscalling due to flanking existing variants	15	12	136	MMAA - c.593_596delCTGA	NA
	Skewed HWE	HWE, $P < 0.001$	93	63	3,126	TTPA - c.744del1	EVS
	Individual data not accessible	Candidates from ESP, individual genotype review and confirmation are not accessible	33	22	88	ALMS1 - c.10769delC	EVS
Inaccurate database information	Polymorphism	Allele frequency too high >0.5%	15	12	6,718	CPT1A - c.1436C>T	Ref. 43
	Pseudodeficiency allele	Pseudodeficiency alleles were defined as DM variant	2	1	1,821	ARSA - c.1055A>G	Ref. 44
	Variant of unknown significance	Published evidence cannot support pathogenicity	3	3	4	CFTR - c.3717+45G>A	Ref. 45
Mutation clinical review	Penetrance	Asymptomatic homozygous carriers were seen	6	6	123	IVD - c.941C>T	Ref. 46
	Age at onset	Homozygotes may show symptoms at adulthood	3	3	20	CNGB3 - c.1208G>A	Refs. 47,48
	Severity	Variable expressivity, homozygotes may present a mild end of phenotype spectrum without drawing medical attention	20	18	90	COL1A1 - c.3897C>G	Ref. 49
	Environmental factor	Disease presentation can be corrected by food avoidance	5	2	11	PAH - c.1241A>G	Ref. 50
	Insufficient evidence	Reasons other than above, like only single case reported	5	5	13	SIL1 - c.274C>T	Ref. 51
Individual clinical review	Cannot pass clinical QC	Expected phenotypes presented	10	9	14	MECP2 - c.1072G>A	RettBase[52]
	Genotype cannot be confirmed	Sanger sequencing shows heterozygous call	11	11	15		NA

NA, not available; EVS, Exome Variant Server (http://evs.gs.washington.edu/EVS/).

Table 4 13 Candidates identified in the Resilience Project

Phenotype	Gene	Mutation (cDNA; protein (reference))	Genomic coordinate (hg19)	Mutation severity	Candidate confidence	Panel source	No. of candidates	Zygosity	Data source	Level of support for candidacy[a]	Sample status	Population carrier frequency[b]	
												1KG	ESP
Cystic fibrosis	*CFTR*	c.1558G>T; p.V520F (NM_000492.3)	Chr7 117199683	Severe pulmonary disease, childhood-onset	Strong	Core allele panel	3	hom	23andMe	C1,C2,C3, G1,G2,G3	2 adults, one declared no manifestation	0.00	0.00
Smith-Lemli-Opitz syndrome	*DHCR7*	c.964-1G>C (NM_001360.2)	Chr11: 71146886	Severe developmental disorder, probably embryonic lethal	Strong	Core allele panel	2	hom	UK10K	C1,C2, G1,G2	Not obtained	0.0052	0.011
Familial dysautonomia	*IKBKAP*	c.2204+6T>C (NM_003640.3)	Chr9: 111662096	Severe neurological disease, high mortality in early childhood	Strong	Core allele panel	1	hom	23andMe	C1,C2, G1,G2,G3	No disease reported by individual	0.00	0.0012 (only in EA)
Epidermolysis Bullosa simplex	*KRT14*	c.373C>T; p.R125C (NM_000526.4)	Chr17: 39742714	Severe dermatologic condition, infantile onset	Strong	Core allele panel	1	het	BGI	C1,C2,C3, G1,G2	No disease reported by individual	0.00	0.00
Pfeiffer syndrome	*FGFR1*	c.755C>G; p.P252R (NM_023110.2)	Chr8: 38282208	Severe congenital skeletal dysplasia with variable expressivity	Strong[c]	Core allele panel	1	het	SWE-SCZ	C1,C2,C3, G1,G2,G3	No abnormal morphology reported in discharged health information	0.00	0.00
APECED	*AIRE*	c.769C>T; p.R257* (NM_000383.2)	Chr21: 45709656	Severe childhood-onset autoimmune disease	Strong	Core allele panel	1	hom	23andMe	C1,C2,C3, G1,G2	No disease reported by individual	0.00	0.00015
Acampomelic campomelic dysplasia	*SOX9*	c.1320C>G; p.Y440* (NM_000346.3)	Chr17: 70120318	Severe skeletal dysplasia with early childhood death	Strong	Expanded panel	1	het	FINN	C1,C2, G1,G2	Not obtained	0.00	0.00
Atelostogenesis	*SLC26A2*	c.835C>T; p.R279W (NM_000112.3)	Chr5: 149359991	Severe early-onset skeletal dysplasia with variable expressivity	Moderate[d]	Expanded panel	3	hom	23andMe	C1,C2, G1,G2	Not obtained	0.0028	0.0023

[a]See **Table 5** for code definitions. [b]Carrier frequencies from combined ethnicities. [c]Individual was categorized as strong candidate due to lack of dysmorphic features. [d]Individual with variable phenotypes have been reported with the mutation[37].
EA, European American.

Attempted recontact of candidate resilient individuals

We were unable to recontact any of the 13 candidate resilient individuals identified in this study, often due to the absence of a recontact clause in the original informed consent forms used for the studies from which these individuals were identified. Although recontact was possible for some cohorts in this study (e.g., Mount Sinai School of Medicine Biobank), no candidates were identified from those cohorts. Given this, we were unable to perform additional critical preprocessing steps to further confirm the resilient status of these individuals. Such steps would include confirming that the analyzed DNA matched the correct medical records for each individual, that they had not been diagnosed with the indicated Mendelian disorder, and that they were not mosaics. We consider these preprocessing steps as critical in order to formally characterize candidates as truly resilient.

Searching for simple explanations of resilience

Although in-depth decoding of candidate resilient individuals requires unfettered access to the individual and their medical records, we searched for counterbalancing variants occurring in the same gene region as the pathogenic one in an attempt to uncover simple explanations for the putative resilience. Among the 13 candidates we identified, 2 from the UK10K cohort had WES data (**Table 4**) and both had the pathogenic variant in the *DHCR7* gene. These two individuals had 14 and 17 additional DHCR7 variants, respectively. Only five of these variants were annotated in the ClinVar, HGMD, and/or OMIM databases (**Supplementary Table 4**). All five were annotated as benign by ClinVar. Interestingly, both of these resilient candidates share the same homozygous alternative genotypes across all five variants. None of the variants identified clearly explains putative resilience in these two individuals. The pathogenic variant in these two individuals alters the splice site acceptor for the last exon (c.964-1G>C). Therefore, in explaining the resilience to this mutation, WGS data would provide a way to search for variants that could lead to the last exon being retained. For the remaining 11 candidates, either the raw sequencing data were inaccessible or only genotype data were available. In these cases the interrogated sites in the implicated gene regions were too sparsely covered to draw conclusions.

Lowering filtering stringency to retrieve more candidates

Given the small number of resilient candidates identified using our high-stringency filters, we attempted to lower their stringency to expand our search. Specifically, we broadened the disease and allele selection criteria to include conditions with more variable or milder clinical manifestations, reduced (but still very high) penetrance, phenotypes that can be managed, and a lower evidence level. These criteria resulted in the identification of 111 additional, second-tier candidates (**Supplementary Table 5**). However, the larger number of candidates resulted in a dramatic increase in the complexity of evaluating their legitimacy compared to that of the first-tier candidates. For example, 33 candidates were associated with conditions with known incomplete penetrance or milder clinical manifestations, 43 harbored variants that were more likely to be polymorphic based on evidence available in the genome variation databases, 7 harbored variants that have been reported only once or in a limited number of patients from the literature, and the remaining 28 candidates had mutations associated with conditions that are known to be strongly influenced by environmental factors. The number of candidates identified were still not large enough to employ statistical genetics techniques to identify modifier loci, and the complexity of the genetic variance component may be significantly increased, making it more challenging to employ

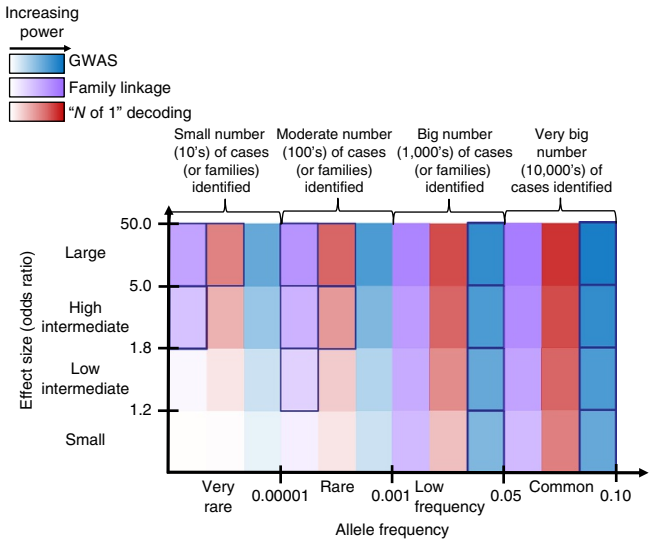

Figure 2 Different strategies for identifying genetic variants buffering human disease. Just as for human diseases, alleles that offer protection against disease can have a broad range of effect sizes and allele frequencies. We depict in a qualitative way the power across the allele frequency and effect size dimensions for three genetic strategies that could be used to identify protective loci: (i) genome-wide association studies (GWAS), (ii) family linkage studies and (iii) "*N* of 1" decoding strategies. For common and low frequency variants, the "*N* of 1" strategy morphs into the statistical GWAS strategy, leveraging the power with adequate numbers that can exist to detect associations between locus genotypes and phenotypes. As allele frequencies decrease, the effect size plays a more crucial role in determining what genetic strategies may be effective for identifying protective alleles (dark blue borders indicate the preferred strategy at the indicated allele frequency and effect size). When the allele frequency is very rare and the effect size is small, there is no effective genetic strategy for identifying such loci, so that other experimental strategies must be employed. However, in the case where very rare, large effect size protective alleles exist, targeted families or "*N* of 1" decoding strategies that depend not on statistical power for detecting associations, but on advanced technologies (genome editing, stem cell reprogramming, DNA/RNA sequencing, computational biology algorithms and so on) combined with appropriate experimentation to elucidate the complexity of protective effects.

variant-specific, or even individual-specific, study designs to elucidate the complexity of resilience (**Fig. 2**).

DISCUSSION

The primary objective of this study was to construct a screening panel to identify individuals who did not have clinical manifestations of severe childhood-onset diseases despite harboring causal mutations believed to be completely penetrant. The multi-tier panel design was driven by technological limitations regarding the characterization of disease mutations, a desire to allow for customization of a screening panel, and by financial considerations in carrying forward a prospective screen for resilient individuals. Although WGS/WES of all participants in such a study would theoretically maximize coverage of genetic information, the associated cost ($300–$1,500/sample) would greatly reduce the number of individuals that could be screened by a targeted sequencing panel (<$50/sample).

The utility of a high-impact screening panel depends directly on rigorous informatics processes and clinical review. Less than 1% of the candidates we initially identified from the screening panel survived our filtering criteria. More than 75% of the initial candidates identified were filtered out due to errors in variant calls resulting from low coverage that made it difficult to reliably call homozygous genotypes, high GC or AT content known to lead to higher sequencing-error rates, or from repetitive sequences known to lead to alignment errors that in turn lead to false small insertion or deletion calls. The remaining false positives represented candidates that failed to pass our established clinical presentation criteria, harbored mutations that were inaccurately represented in the mutation databases, or for which there was insufficient scientific evidence to support the predicted phenotypic impact of the mutation.

Table 5 Status codes for different levels of support identified during follow up of candidate resilient individuals

Support type	Status code	Status description for different levels of support for candidacy
Clinical validation	C1	Pass criteria for severity and penetrance for specific mutation set and reviewed by clinical specialist
	C2	Reference in literature found that can be cited for that mutation
	C3	Individual's clinical record examined - lacking classical presentation by "chart review" and family history
	C4	Individual is able to be recontacted to confirm atypical clinical presentation
Genetic validation	G1	Genotype call made
	G2	Review of primary sequencing/genotyping data
	G3	Resequencing of the sample
	G4	Work-up to rule out mosaic
Biomedical validation	B	Clinical test performed to determine if the individual harbor expected biomedical characteristics (enzyme activity, blood count, organ function etc.)

Of the identified candidate resilient individuals, two individuals from the UK10K project were homozygous carriers of a splicing consensus acceptor mutation for Smith-Lemli-Opitz syndrome (SLOS). This is a well-known mutation leading to a null allele of the delta-7-sterol reductase gene, which accounts for up to one-third of mutant alleles of SLOS patients in populations of European descent. Homozygotes of this splicing mutation are rarely seen in SLOS patients despite the high carrier frequency, and all manifest at the severe end of the SLOS phenotypic spectrum and are not known to survive through childhood[26,27]. Four other well-characterized recessive diseases were represented in our final list of candidates. The *CFTR* mutation c.1558G>T is associated with classic cystic fibrosis in combination with other disease alleles, but no homozygous cases have been described to the best of our knowledge. *In vitro* analysis has demonstrated that the mutated form of the CTFR receptor traffics to the cell surface but has severely impaired function[28]. The *IKBKAP* mutation is an Ashkenazi Jewish founder mutation observed in nearly all cases of familial dysautonomia, a debilitating childhood-onset disorder[29]. The Finnish/European c.769C>T mutation in *AIRE* has been associated with autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy syndrome (APECED)[30], a childhood-onset disorder characterized by chronic mucocutaneous candidiasis, hypoparathyroidism and Addison's disease. The p.R279W is a common *SLC26A2* mutation. Compound heterozygotes or homozygotes of this mutation usually manifest severe skeletal dysplasia, although patients with milder phenotypes have been reported[31].

Three autosomal dominant disorders are represented in our final list of candidates. The *KRT14* c.373C>T mutation has been associated with the severe Dowling-Meara subtype of epidermolysis bullosa simplex (MIM131760)[32]. The recurrent c.755C→G mutation

in *FGFR1* has been associated with Pfeiffer syndrome, a craniosynostosis disorder with manifestations in the distal extremities[33]. The *SOX9* nonsense mutation p.Y440* is recurrently seen in patients with acampomelic campomelic dysplasia (MIM114290)[34–36], a severe form of skeletal dysplasia. Variable survival time of patients with this same mutation and lack of clear genotype-phenotype correlation among patients suggest that genetic modifiers that affect phenotypic variability may exist.

During our screening of the existing data sets, we identified a *GBA* compound-heterozygous (affecting amino acid positions p.N409S and p.L483P in the protein sequence) individual who had undergone routine carrier screening at Mount Sinai, but who had never been diagnosed with Gaucher disease. Upon clinical review, it was demonstrated that this individual exhibited subclinical manifestations of this disease. This patient's diagnosis was subsequently confirmed by acid β-glucosidase assay, which was in the affected range (0.7 nmol/h/mg, range 3.6–18.2 nmol/h/mg). Her medical record showed a history of easy bruising and bleeding since childhood; she was subsequently misdiagnosed with idiopathic thrombocytopenic purpura. The patient currently receives enzyme replacement therapy, which has resulted in improvement with respect to thrombocytopenia. Her story is an example of the complexity of genetic conditions such as Gaucher disease, which can exhibit a broad range of expressivity, leading to subclinical manifestations and misdiagnoses.

Given that most of the candidate resilient individuals were unavailable for recontacting, we cannot exclude straightforward explanations for their candidacy status. With the exception of disorders with hematologic manifestations, somatic mosaicism for deleterious mutations could explain the absence of phenotypic expression. The 589,306 individuals analyzed in this study were recruited from 12 large study cohorts, where the sample types were mixed with respect to ethnicity and health status, providing for the possibility that one or more of the candidates in our final list was an affected individual that harbors a homozygous deleterious mutation that may explain their diagnosed condition. The lack of metadata and the unavailability for recontacting of those participating in this study present perhaps the biggest obstacles for leveraging data retrospectively to identify resilient individuals, and speaks to the advantage of carrying out a prospective search for resilient individuals where participants can be appropriately consented for recontacting, and relevant metadata can be collected.

Despite the difficulties in getting traction on decoding the 13 individuals we identified, a number of findings demonstrate the utility of carrying out this type of comprehensive screen. First, we found mutations for severe early-onset diseases that are annotated as being completely penetrant, in putative nonpenetrant individuals, providing for the possibility that genetic modifiers may be more common than believed. Therefore, identification of resilient individuals may enhance our understanding of Mendelian disease etiology and how we counsel others regarding such conditions. Second, our screening panel provides a fully curated list of variants and their disease implications that go beyond what is covered by currently available commercial screening panels. Finally, our study suggests that genotype calling and disease variant curation and annotation are still a challenge for deriving meaningful interpretations from large-scale genomic data.

The extremely rare frequency of candidate resilient individuals in this retrospective study supports the intuitive notion that securing larger numbers of candidates would require analyzing all data worldwide being generated by genotyping and next-generation sequencing methods. A number of existing projects, such as the Human Knockout Project[37], The Million Veterans Program[38] and the large UK Biobank Project[39], all stand to contribute considerably to this type of effort. Whereas the penetrance, disease severity and allele-frequency parameters employed in our study restricted our screen to those mutations thought to be completely penetrant with very severe childhood manifestations of disease phenotypes, a broader net could be cast by relaxing these conditions, and allowing, for example, mutations that are not completely penetrant, but still highly penetrant (**Fig. 2**). Although this would result in an increase in the number of candidate resilient individuals, it would come at the expense of increasing the complexity of the factors buffering disease. We observed a sharp increase in the number of candidates by slightly loosening our stringency filters (**Supplementary Table 5**), but this increase was complemented by an increase in the complexity of interpretation, annotation and subsequent follow-up analyses for these additional candidates. It is worth trying to understand the complex tradeoffs between sample size, penetrance, the genetic complexity of the disease as well as resilience to disease, and our ability to identify factors buffering the disease (**Fig. 2**).

In prospective searches for resilient individuals, more appropriate consenting will be needed to link participants to their medical records and to allow for appropriate recontacting that enables follow-up characterizations, validation of their resilient condition and decoding to uncover the causes of the resilience. In cases where the buffering effect is itself a highly penetrant Mendelian trait, even with a small sample size (even a sample size of 1, referred to as "*N* of 1" cases), there is a reasonable probability of identifying the genetic cause. For example, a number of studies using whole-exome sequencing to provide diagnoses for undiagnosed, suspected genetic conditions, resulted in a roughly 25% success rate, with a significant proportion of these successes resulting in the identification of mutations that had not been previously characterized[40]. In "*N* of 1" cancer cases for both retrospective[41] and prospective studies[42], finding actionable mutations that can affect treatment choices happens in well over 50% of the cases, with a high percentage of the actionable mutations identified as being *de novo*. We anticipate that future searches for individuals resilient to various genetic defects will be most effective when combining the traditional searches for positive outliers in known extended families with very broad searches for positive outliers in the general population.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Sieberts (Sage Bionetworks) and L. Mangravite (Sage Bionetworks) for critical review of our manuscript. The authors would like to thank the Exome Aggregation Consortium and the group that provided exome variant data for comparison.

AUTHOR CONTRIBUTIONS

R.C., E.E.S. and S.H.F. contributed to the conception and study design. L.S. and R.C. built the disease gene and mutation panels. J.H. and R.C. curated databases and built bioinformatics pipelines. J.H. and R.C. performed bioinformatics analysis. L.S., R.C., J.H., B.N., M.A.D., E.Z., G.A.D., L.E. and S.H.F. performed QC and clinical review of all candidates. B.N., P. Sklar, J.Z., H.Z., L.T., O.P., M.L., P. Sleiman, W.-y.C., W.C., H.S., Y.S., M.F., L.O., J.R.B., E.L., T.H., L.G., J.W., H.H. and A.W. contributed the data and analysis. L.S., R.C., J.H., E.E.S. and S.H.F. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
2. Stenson, P.D. *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).
3. Dietz, H.C. New therapeutic approaches to mendelian disorders. *N. Engl. J. Med.* **363**, 852–863 (2010).
4. Topol, E.J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241–253 (2014).
5. Bell, C.J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
6. Lazarin, G.A. *et al.* An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals. *Genet. Med.* **15**, 178–186 (2013).
7. Tanner, A.K. *et al.* Development and performance of a comprehensive targeted sequencing assay for pan-ethnic screening of carrier status. *J. Mol. Diagn.* **16**, 350–360 (2014).
8. Hartman, J.L. IV. Buffering of deoxyribonucleotide pool homeostasis by threonine metabolism. *Proc. Natl. Acad. Sci. USA* **104**, 11700–11705 (2007).
9. Hartman, J.L. IV., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291**, 1001–1004 (2001).
10. Hartman, J.L. IV. & Tippery, N.P. Systematic quantification of gene interactions by phenotypic array analysis. *Genome Biol.* **5**, R49 (2004).
11. Louie, R.J. *et al.* A yeast phenomic model for the gene interaction network modulating CFTR-ΔF508 protein biogenesis. *Genome Med.* **4**, 103 (2012).
12. Philpott, S. *et al.* CCR5 genotype and resistance to vertical transmission of HIV-1. *J. Acquir. Immune Defic. Syndr.* **21**, 189–193 (1999).
13. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
14. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
15. Pasmooij, A.M. *et al.* Revertant mosaicism due to a second-site mutation in COL7A1 in a patient with recessive dystrophic epidermolysis bullosa. *J. Invest. Dermatol.* **130**, 2407–2411 (2010).
16. Ikeda, H. *et al.* Genetic reversion in an acute myelogenous leukemia cell line from a Fanconi anemia patient with biallelic mutations in BRCA2. *Cancer Res.* **63**, 2688–2694 (2003).
17. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
18. Vieira, N.M. *et al.* Jagged 1 rescues the Duchenne muscular dystrophy phenotype. *Cell* **163**, 1204–1213 (2015).
19. Friend, S.H. & Schadt, E.E. Translational genomics. Clues from the resilient. *Science* **344**, 970–972 (2014).
20. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
21. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
22. Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur. J. Hum. Genet.* **22**, 1100–1104 (2014).
23. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* Preprint at http://biorxiv.org/content/early/2015/10/30/030338 (2015).
25. Cheng, W.Y., Hakenberg, J., Li, S.D. & Chen, R. DIVAS: a centralized genetic variant repository representing 150 000 individuals from multiple disease cohorts. *Bioinformatics* **32**, 151–153 (2016).
26. Jira, P.E. *et al.* Novel mutations in the 7-dehydrocholesterol reductase gene of 13 patients with Smith–Lemli–Opitz syndrome. *Ann. Hum. Genet.* **65**, 229–236 (2001).
27. Nowaczyk, M.J. *et al.* Smith-Lemli-Opitz (RHS) syndrome: holoprosencephaly and homozygous IVS8-1G-->C genotype. *Am. J. Med. Genet.* **103**, 75–80 (2001).
28. Sosnay, P.R. *et al.* Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet.* **45**, 1160–1167 (2013).
29. Shohat, M. & Hubshman, M.W. in *GeneReviews* (eds. Pagon, R.A. *et al.*) (University of Washington, Seattle, 1993–2016) (updated December 18, 2014).
30. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nat. Genet.* **17**, 393–398 (1997).
31. Ballhausen, D. *et al.* Recessive multiple epiphyseal dysplasia (rMED): phenotype delineation in eighteen homozygotes for DTDST mutation R279W. *J. Med. Genet.* **40**, 65–71 (2003).
32. Letai, A. *et al.* Disease severity correlates with position of keratin point mutations in patients with epidermolysis bullosa simplex. *Proc. Natl. Acad. Sci. USA* **90**, 3197–3201 (1993).
33. Muenke, M. *et al.* A common mutation in the fibroblast growth factor receptor 1 gene in Pfeiffer syndrome. *Nat. Genet.* **8**, 269–274 (1994).
34. Ebensperger, C. *et al.* No evidence of mutations in four candidate genes for male sex determination/differentiation in sex-reversed XY females with campomelic dysplasia. *Ann. Genet.* **34**, 233–238 (1991).
35. Wagner, T. *et al.* Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. *Cell* **79**, 1111–1120 (1994).
36. Meyer, J. *et al.* Mutational analysis of the SOX9 gene in campomelic dysplasia and autosomal sex reversal: lack of genotype/phenotype correlations. *Hum. Mol. Genet.* **6**, 91–98 (1997).
37. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
38. Roberts, J.P. Million veterans sequenced. *Nat. Biotechnol.* **31**, 470 (2013).
39. Palmer, L.J.U.K. UK Biobank: bank on it. *Lancet* **369**, 1980–1982 (2007).
40. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *J. Am. Med. Assoc.* **312**, 1880–1887 (2014).
41. Schwaederle, M. *et al.* On the road to precision cancer medicine: analysis of genomic biomarker actionability in 439 patients. *Mol. Cancer Ther.* **14**, 1488–1494 (2015).
42. Beltran, H. *et al.* Whole-exome sequencing of metastatic cancer and biomarkers of treatment response. *JAMA Oncol.* **1**, 466–474 (2015).
43. Rajakumar, C. *et al.* Carnitine palmitoyltransferase IA polymorphism P479L is common in Greenland Inuit and is associated with elevated plasma apolipoprotein A-I. *J. Lipid Res.* **50**, 1223–1228 (2009).
44. Fluharty, A.L. in *GeneReviews* (eds. Pagon, R.A. *et al.*) (University of Washington, Seattle, 1993–2016) (updated February 6, 2014).
45. Bienvenu, T. *et al.* Spectrum of CFTR mutations on Réunion Island: impact on neonatal screening. *Hum. Biol.* **77**, 705–714 (2005).
46. Ensenauer, R. *et al.* A common mutation is associated with a mild, potentially asymptomatic phenotype in patients with isovaleric acidemia diagnosed by newborn screening. *Am. J. Hum. Genet.* **75**, 1136–1142 (2004).
47. Samstad, S.O., Rossvoll, O., Torp, H.G., Skjaerpe, T. & Hatle, L. Cross-sectional early mitral flow-velocity profiles from color Doppler in patients with mitral valve disease. *Circulation* **86**, 748–755 (1992).
48. Thiadens, A.A. *et al.* Comprehensive analysis of the achromatopsia genes CNGA3 and CNGB3 in progressive cone dystrophy. *Ophthalmology* **117**, 825–30.e1 (2010).
49. Pace, J.M., Kuslich, C.D., Willing, M.C. & Byers, P.H. Disruption of one intra-chain disulphide bond in the carboxyl-terminal propeptide of the proalpha1(I) chain of type I procollagen permits slow assembly and secretion of overmodified, but stable procollagen trimers and results in mild osteogenesis imperfecta. *J. Med. Genet.* **38**, 443–449 (2001).
50. Okano, Y. *et al.* Molecular basis of phenotypic heterogeneity in phenylketonuria. *N. Engl. J. Med.* **324**, 1232–1238 (1991).
51. Riazuddin, S.A. *et al.* Novel SIL1 mutations in consanguineous Pakistani families mapping to chromosomes 5q31. *Mol. Vis.* **15**, 1050–1056 (2009).
52. Christodoulou, J., Grimm, A., Maher, T. & Bennetts, B. RettBASE: The IRSA MECP2 variation database-a new mutation database in evolution. *Hum. Mutat.* **21**, 466–472 (2003).

ONLINE METHODS

Curating a mutation database of severe childhood Mendelian disorders. The first step in our workflow for interrogating existing large-scale sequence and genotype data (**Supplementary Fig. 1**) is the construction of a comprehensive gene panel comprising genes that harbor completely penetrant mutations for severe childhood Mendelian disorders. We consolidated gene and mutation information for such disorders from eight independent databases that contained complementary and supporting data for genes and mutations involved in disease: (i) the Online Mendelian Inheritance in Man (OMIM) database (http://www.omim.org/)[1]; (ii) the Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk)[2]; (iii) GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/)[18]; (iv) Genetics Home Reference (GHR; http://ghr.nlm.nih.gov/); (v) ClinVar (http://www.clinvar.com/)[53]; (vi) Orphanet (http://www.orpha.net)[54]; (vii) the Leiden Open Variation Database (LOVD; http://www.lovd.nl/3.0/home)[55]; and (viii) Reference Variant Store (RVS)[56].

Criteria for including diseases and alleles in our database. To restrict attention to severe childhood Mendelian disorders, we required a disease to have certain features to be represented on our panel. First, we required the disease to be a Mendelian disorder with known pathogenic mutation(s) and a clear mode of inheritance: autosomal recessive, autosomal dominant or X-linked recessive. Disorders arising from mitochondrial DNA variants or the many different types of structural variants, digenic and complex diseases were not considered. Second, we restricted our attention to diseases that were not exceptionally rare, defined as having a prevalence higher than one in one million individuals or an increased incidence in specific subpopulations. Third, we restricted attention to diseases in which patients manifest severe, obvious phenotypes that lead to significantly increased mortality or are debilitating early in life. Fourth, we required that the clinical manifestation of the disease most typically occur before 18 years of age. Finally, we required that the diseases be caused by (nearly) completely penetrant mutations (**Supplementary Table 6** and **Supplementary Fig. 6**).

For the set of diseases represented in our screening panels, there may be many mutations that can cause them, but the expressivity of these mutations can vary widely with respect to age of onset, severity and penetrance. We focused on those mutations that were completely penetrant and that led to the most severe forms of disease. Therefore, we constructed a filter that ensured the mutations on our panel met these different criteria. First, we required the mutation to be recurrent (a 'hotspot'), seen in multiple patients or reported several times in literature, or that it be a known founder mutation in a given subpopulation. Second, we required that the mutation be fully penetrant or nearly completely penetrant. Third, we required the mutations to be associated with severe phenotypes, having significantly increased mortality or debilitation before adulthood. Fourth, we required that the mutations lead to a significant loss of production or function compared to normal mRNAs or proteins (nonsense mutations, frameshift mutations that lead to premature stop codons or missense mutations known to affect important protein domains). Finally, we restricted attention to those mutations that could be more easily detected by standard genotyping or sequencing assays. Mutations that involve gross genomic rearrangement, copy number abnormality, large deletion/insertion and tandem repeats, although highly interesting, were excluded from consideration given that the DNA variant information available for our study did not include these types of calls and most of the data used in this study were generated by technologies and protocols that were not optimized to routinely assay structural variants in a high-throughput fashion. For example, more than half of the samples examined in this study relied on existing genotype data sets from which these types of mutations cannot be reliably called.

Deriving a screening panel to identify individuals resilient to severe childhood Mendelian disorder. From the set of rare Mendelian childhood diseases, genes and associated mutations assembled above, we derived a gene panel and two allele panels to employ in our screen. The gene panel comprised curated genes associated with early-onset severe disease, and the two allele panels comprised disease-causing mutations that were identified at different confidence levels. For the gene panel, we compiled a list of genes associated with the highly penetrant, early-onset, severe Mendelian disorders identified above. The clinical significance for the diseases and corresponding mutations was annotated based on information from public human genetics disease phenotype databases (OMIM, GeneReviews, Genetic Testing Repository, GHR, ClinVar, Orphanet), the literature and published carrier-screening panels[5–7] (**Supplementary Fig. 7a**). We also used a pre-existing in-house (maintained by R.C.) set of more than 20,000 full-text articles curated for risk alleles and gene-disease associations. Each disease and the corresponding genes harboring mutations were annotated using published data on mode of inheritance, severity, penetrance, prevalence and age of onset. We grouped annotations for each of these annotation types into discrete categories to enable more efficient sorting and filtering (**Supplementary Table 7**). For example, "age of onset" ranges from 1 (prenatal or congenital or infantile <2 years old) to 4 (late onset >18 years old), and then 5 indicating the age of onset is unknown.

The two allele panels were developed from the same sources but using different stringencies. The first panel, CAP, contained only recurrent or founder mutations that had been well-documented and were associated with the most severe phenotype as represented in the above gene panel. Genotype-phenotype correlations and recurrence of mutations were determined based upon the genomic phenotype databases, including OMIM, GeneReviews, ClinVar and LOVD. The CAP was also annotated with respect to a mutation-based clinical significance score assigned to each variant using the same scoring system indicated above (**Supplementary Table 7**). The CAP comprised only the most heavily curated, highest-confidence alleles that are well-established as causing severe childhood disorders. Most of the alleles in the CAP are routinely assayed on carrier screening panels. However, to better leverage the vast number of discoveries made in the last couple of decades, we constructed a second "expanded allele panel" (EAP) that included all disease-associated variants in HGMD classified as disease causing, "DM", and with overall minor allele frequency (MAF) < 0.5% according to the 1000 Genomes and ESP databases, for those genes contained within the gene panel defined above. The rationale for the EAP in addition to CAP was to broaden coverage by leveraging the extensive HGMD resource, accepting the increased noise present in this database for the initial screen, then applying more in-depth curation and clinical review to those variants in the EAP identified as hits. In this way, the significant informatics and clinical resources needed to curate disease alleles were restricted to those identified in our study population. The CAP overlaps significantly with the EAP, but given the extensive curation of the CAP, there are alleles in CAP not represented in EAP (**Supplementary Fig. 7b**). Both allele panels include variant-specific information such as genomic coordination; dbSNP rs-number; cDNA and protein level change in Human Genome Variation Society nomenclature[57], literature references; and most importantly, observation frequencies obtained from several public databases such as 1000 Genomes, ESP6500 and TCGA (normal samples).

Samples analyzed in the Resilience Project. All study subjects in the current retrospective study were from 12 past and ongoing genetic studies worldwide (**Table 2**). Many of these studies provide open, unrestricted access or restricted access through data access committees to the genetic variant data generated in the study, including the 1000 Genomes Project[20], ESP[21], matched normal samples from The Cancer Genome Atlas (TCGA) Project, the UK10K project[22], the SWE-SCZ exome sequencing project, and SISu, whereas others represent private databases that are available through collaboration with the corresponding investigator, such as the Finnish study cohort (which includes the FINRISK cohort, EUFAM, the Finnish Twin Study and the Migraine Study), the Mount Sinai BioBank, 23andMe, BGI exome sequencing database and the Children's Hospital of Philadelphia (CHOP) BioBank.

A wide variety of assays were leveraged in these different studies to score DNA variants, from genotyping of comprehensive SNP panels capturing all common small-nucleotide variation in the genome, to whole exome and genome sequencing (**Table 2**). For imputation of genotyping data sets (Mount Sinai BioBank and CHOP), we used 1,000 Genome Project Phase 1 (b37) as the reference panel. For other genotyping data sets (23andMe and FINN), original assayed genotypes were used. A total of 589,306 individuals' variant data sets were analyzed, including 518,721 genotyping data sets and 70,585 whole exome or whole genome sequencing data sets.

The search for resilient individuals. The union of the CAP and EAP were input into a software tool, Search Your Genome, we developed to screen

genotype and sequence data for disease-causing alleles. Our scanning tool takes Variant Call Format (VCF) files as well as GFF and tab-delimited files, stored either as data summarized across a study or as single sample data sets. The input files were preprocessed by compressing and indexing them using SAMtools bgzip and tabix, respectively[58], with preliminary annotations assigned using snpEff[59] for genes (HGNC symbol or Entrez Gene ID) and nucleotide changes for variants. For VCF files, a set of common markups referring to features such as genotypes, allele frequencies and zygosity were identified for each sample and each variant of interest as defined in our panels, in addition to searching for *de novo* variants in genes represented in our panels. For other input formats, depending on the details provided in the corresponding data files, our tool interrogates the files for homozygotes and compound heterozygotes for alleles in the combined CAP and EAP, as well as for *de novo* variants leading to premature stop codons, given such variants are likely to lead to the same effects as the known deleterious mutations represented in our allele panels. The Search Your Genome tool is written in Java to ensure maximum portability to any platform running a Java Virtual Machine version 6.0 or above. On a typical desktop computer, interrogating the 1000 Genomes data (more than 37 million genetic variants) for resilient individuals from the CAP takes roughly one minute. The software is available at https://bitbucket.org/rongchenlab/resilience and http://rongchenlab.org/software/the-resilience-project-software/.

Manual review and annotation of candidates. For each candidate that has passed high-throughput sequencing and/or genotyping QC pipeline, manual review was performed in small batches by two to five reviewers independently. At least one of the reviewers was a specialist in the disease area associated with the candidate's mutation. Any candidate that achieved consistent categorization from different reviewers, went directly to the final candidate table (if it passed clinical QC) or it was removed from CAP/EAP. For any inconsistent annotations, a group meeting session was called, a deep literature review was done and an extensive discussion was held on clinical significance to guarantee that all candidates in the final resilient individual table had solid evidence of being a real candidate. If the group discussion could not achieve a unified categorization for a candidate, this candidate was rejected from the final candidate table.

53. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
54. Rath, A. *et al.* Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).
55. Fokkema, I.F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563 (2011).
56. Hakenberg, J. *et al.* Integrating 400 million variants from 80,000 human samples with extensive annotations: towards a knowledge base to analyze disease cohorts. *BMC Bioinformatics* **17**, 24 (2016).
57. Horaitis, O. & Cotton, R.G. The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum. Mutat.* **23**, 447–452 (2004).
58. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
59. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92 (2012).

doi:10.1038/nbt.3514

Corrigendum: Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases

Rong Chen, Lisong Shi, Jörg Hakenberg, Brian Naughton, Pamela Sklar, Jianguo Zhang, Hanlin Zhou, Lifeng Tian, Om Prakash, Mathieu Lemire, Patrick Sleiman, Wei-yi Cheng, Wanting Chen, Hardik Shah, Yulan Shen, Menachem Fromer, Larsson Omberg, Matthew A Deardorff, Elaine Zackai, Jason R Bobe, Elissa Levin, Thomas J Hudson, Leif Groop, Jun Wang, Hakon Hakonarson, Anne Wojcicki, George A Diaz, Lisa Edelmann, Eric E Schadt & Stephen H Friend

Nat. Biotechnol. **doi:10.1038/nbt.3514; corrected online 21 April 2016**

In the version of this article initially published, in Table 3, the row labeled "Individual clinical review," the number of mutations should have read 10, not 6; the number of diseases, 9, not 5; and the number of individuals, 14, not 10. The errors have been corrected for the print, PDF and HTML versions of this article.