Data-driven consideration of genetic disorders for global genomic newborn screening programs

Thomas Minten, Sarah Bick, Sophia Adelson, Nils Gehlenborg, Laura M. Amendola, François Boemer, Alison J. Coffey, Nicolas Encina, Alessandra Ferlini, Janbernd Kirschner, Bianca E. Russell, Laurent Servais, Kristen L. Sund, Ryan J. Taft, Petros Tsipouras, Hana Zouk, the ICoNS Gene List Contributors, Robert C. Green, Nina B. Gold, David Bick, Mattia Gentile, Paola Orsini, Romina Ficarella, Maria Luisa Valente, Emanuela Ponzi, Athina Ververi, Maria Koutsogianni, Huang Xinwen, Xiao Rui, Zhao Zhengyan, Matthew J. Pelo, Jovanka King, Carol Siu, Karin Kassahn, Stefaan Sansen, Enrico Bertini, Aldona Zygmunt, the International Consortium on Newborn Sequencing (ICoNS), Sophia Adelson, Mattia Gentile, Mette Nyegaard, Emanuele Agolini, Jessica Giordano, Justin O'Sullivan, Aljazi Al-Maraghi, Ulrich Glumer Jensen, Jelili Ojodu, Karla Alex, David Godler, Paola Orsini, Fowzan Alkuraya, Nina Gold, Andrea Oza, Ammira Alshabeeb Akil, Aaron Goldenberg, Katrina Paleologos, Munira Alshehri, Katie GoldenGrant, Richard Parad, Derek Ansel, Cassie Goldman, Holly Peay, Niki Armstrong, José Manuel González de Aledo-Castillo, Matthew Pelo, Matthew Aujla, Daniel Gottlieb, Carolyn Philstrom, Don Bailey, Robert Green, Dominique Pichard, Mei Baker, Christopher Greene, Amanda Pichini, Jorune Balciuniene, Brooke Greenstein, Holly Pickering, Andrew Barry, Scott Grosse, Michelle Pirreca, Bruce Bennetts, Annette Grueters, Malgorzata Ponikowska, Melissa Berenger, Gulcin Gumus, Amy Ponte, Jonathan Berg, Kelly Hagman, Andreas Posch, Donna Bernstein, Kevin Hall, Cynthia Powell, Arindam Bhatatcharjee, Aymeric Harmant, Liana Protopsaltis, Sucheta Bhatt, Sally Hartmanis, Yeyson Quevedo, David Bick, Robin Hayeems, Marianna Raia, Tracey Bishop, Rose Heald, Rebecca Reimers, Asaf Bitton, Madhuri Hegde, Andy Rohrwasser, François Boemer, Rebecca Heiner-Fokkema, Paul Rollier, Natasha Bonhomme, Lidewij Henneman, Lene Rottensten, George Bowley, Becca Hernan, Irakli Rtskhiladze, Brenna Boyd, Charlotte Hobbs, Nabihah Sachedina, Heiko Brennenstuhl, Ingrid Holm, George Sahyoun, Steven Brenner, Layla Horwitz, Aditi Satija, Mairead Bresnahan, Zhanzhi Hu, Christian Schaaf, Thomas Brewster, Maria Iascone, Jennifer Schleit, P.J. Brooks, Ken Irvine, Richard Scott, Katya Broomberg, Guanjun Jin, Lauren Scully, Amy Brower, Kelsey Kalbfleisch, Stacey Seeloff, Gemma Brown, Ines Kander, Laurent Servais, James Buchanan, Lucy Kaplun, Nidhi Shah, Caleb Bupp, Dalia Kasperaviciute, Maija Siitonen, Candance Cameron, Karin Kassahn, Sikha Singh, Lauren Capacchione, Leni Kauko, Carol Siu, Diana Carli, Riina Kaukonen, Hadley Smith, Onassis Castillo Ceballo, Nicole Kelly, Lisa Sniderman King, Kee Chan, Dhayo Khangsar, Neal Sondheimer, Jillian Chance, Jovanka King, Lourdes St George, Georgia Charalambidou, Clare Kingsley, Zornitza Stark, Winnie Chen, Stephen Kingsmore, Robert Steiner, Yun-Ru Chen, Brian Kirmse, Ulrik Stoltze, Wendy Chung, Rachel Klein, Asbjørg Stray-Pedersen, Brian Chung, Stefan Koelker, Kristen Sund, Megan Clarke, Youssef Kousa, Paris Tafas, Susan Clasper, Elizaveta Krupoderova, Polakit Teekakirikul, F. Sessions Cole, Paul Kruszka, Dimitrios Thanos, Heidi Cope, Katherine Langley, Audrey Thurm, Stephanie Coury, Ciara

Leckie, Meekai To, Tony Cox, Emmanuelle Lecommandeur, Petros Tsipouras, Tamara Dangouloff, David Ledbetter, Alice Tuff-Lacey, Earnest James Paul Daniel, Pamela Lee, Heather Turner, Katrin Eivindardottir Danielsen, Beomhee Lee, Philip Twiss, Emeline Davoine, Camille Level, Fiona Ulph, Tom Defay, Celine Lewis, Daniel Uribe, Geethanjali Devadoss Gandhi, Anna Lewis, Tiina Urv, Joseph Dewulf, Ruby Liu, Cora Vacher, Lisa Diller, Mauro Longoni, Kris Van Den Bogaert, Pakhi Dixit, Alberte Lundquist, Mirjam van der Burg, Martijn Dolle, Sebastian Lunke, Eva Van Steijvoort, Lilian Downie, Kate MacDuffie, Yiota Veloudi, Erin Drake, Ankit Malhotra, Elizabeth Vengoechea, Suzanne Drury, Lionel Marcelis, Els Voorhoeve, Annelotte Duintjer, Maria Martinez-Fresno, Martin Vu, Bugrahan Duz, Gert Matthijs, Melissa Wasserstein, David Eckstein, Roberts Melbardis, Michael Watson, Matthew Ellinwood, Jessica Merritt, Bryn Webb, Katarzyna Ellsworth, Radja Messai Badji, Anna Wedell, Sarah Elsea, Lou Metherell, Thomas Westover, Nicolas Encina, Nanna Balle Mikkelsen, Alexandra Wiedemann, Harriet Etheredge, Laura Milko, Meredith Wright, Laurence Faivre, Nicole Miller, Cindy Wu, Alessandra Ferlini, Thomas Minten, Julie Yeo, Monica Ferrie, Sian Morgan, Nancy Yin-Hsiu Chien, Terri Finkel, Katarzyna Mosiewicz, Shamila Yusuff, Petra Furu, Ulrike Mütze, Tomasz Zemojtel, Jamie Galarza-Cornejo, Sukhvinder Nicklen, Bethany Zettler, Ya Gao, Minna Niemela, Zhengyan Zhao, Judit Garcia-Villoria, Dau-Ming Niu, Joanna Ziff, Liz Gardner, Sarah Norris, Rebekah Zimmerman, Amy Gaviglio, Antonio Novelli, Michela Zuccolo, Michael Gelb, Arwa Nusair

PII: S1098-3600(25)00090-5

DOI: https://doi.org/10.1016/j.gim.2025.101443

Reference: GIM 101443

To appear in: *Genetics in Medicine*

Received Date: 22 December 2024

Revised Date: 8 April 2025

Accepted Date: 11 April 2025

Greene C, Pichini A, Balciuniene J, Greenstein B, Pickering H, Barry A, Grosse S, Pirreca M, Bennetts B, Grueters A, Ponikowska M, Berenger M, Gumus G, Ponte A, Berg J, Hagman K, Posch A, Bernstein D, Hall K, Powell C, Bhatatcharjee A, Harmant A, Protopsaltis L, Bhatt S, Hartmanis S, Quevedo Y, Bick D, Hayeems R, Raia M, Bishop T, Heald R, Reimers R, Bitton A, Hegde M, Rohrwasser A, Boemer F, Heiner-Fokkema R, Rollier P, Bonhomme N, Henneman L, Rottensten L, Bowley G, Hernan B, Rtskhiladze I, Boyd B, Hobbs C, Sachedina N, Brennenstuhl H, Holm I, Sahyoun G, Brenner S, Horwitz L, Satija A, Bresnahan M, Hu Z, Schaaf C, Brewster T, Iascone M, Schleit J, Brooks P, Irvine K, Scott R, Broomberg K, Jin G, Scully L, Brower A, Kalbfleisch K, Seeloff S, Brown G, Kander I, Servais L, Buchanan J, Kaplun L, Shah N, Bupp C, Kasperaviciute D, Siitonen M, Cameron C, Kassahn K, Singh S, Capacchione L, Kauko L, Siu C, Carli D, Kaukonen R, Smith H, Ceballo OC, Kelly N, King LS, Chan K, Khangsar D, Sondheimer N, Chance J, King J, St George L, Charalambidou G, Kingsley C, Stark Z, Chen W, Kingsmore S, Steiner R, Chen YR, Kirmse B, Stoltze U, Chung W, Klein R, Stray-Pedersen A, Chung B, Koelker S, Sund K, Clarke M, Kousa Y, Tafas P, Clasper S, Krupoderova E, Teekakirikul P, Cole FS, Kruszka P, Thanos D, Cope H, Langley K, Thurm A, Coury S, Leckie C, To M, Cox T, Lecommandeur E, Tsipouras P, Dangouloff T, Ledbetter D, Tuff-Lacey A, Paul Daniel EJ, Lee P, Turner H, Danielsen KE, Lee B, Twiss P, Davoine E, Level C, Ulph F, Defay T, Lewis C, Uribe D, Gandhi GD, Lewis A, Urv T, Dewulf J, Liu R, Vacher C, Diller L, Longoni M, Van Den Bogaert K, Dixit P, Lundquist A, van der Burg M, Dolle M, Lunke S, Van Steijvoort E, Downie L, MacDuffie K, Veloudi Y, Drake E, Malhotra A, Vengoechea E, Drury S, Marcelis L, Voorhoeve E, Duintjer A, Martinez-Fresno M, Vu M, Duz B, Matthijs G, Wasserstein M, Eckstein D, Melbardis R, Watson M, Ellinwood M, Merritt J, Webb B, Ellsworth K, Badji RM, Wedell A, Elsea S, Metherell L, Westover T, Encina N, Mikkelsen NB, Wiedemann A, Etheredge H, Milko L, Wright M, Faivre L, Miller N, Wu C, Ferlini A, Minten T, Yeo J, Ferrie M, Morgan S, Yin-Hsiu Chien N, Finkel T, Mosiewicz K, Yusuff S, Furu P, Mütze U, Zemojtel T, Galarza-Cornejo J, Nicklen S, Zettler B, Gao Y, Niemela M, Zhao Z, Garcia-Villoria J, Niu DM, Ziff J, Gardner L, Norris S, Zimmerman R, Gaviglio A, Novelli A, Zuccolo M, Gelb M, Nusair A, Data-driven consideration of genetic disorders for global genomic newborn screening programs, *Genetics in Medicine* (2025), doi: https://doi.org/10.1016/j.gim.2025.101443.

Data-driven consideration of genetic disorders for global genomic newborn screening programs

Thomas Minten[1], Sarah Bick[2,3,4], Sophia Adelson[5,6], Nils Gehlenborg[7], Laura M. Amendola[8], François Boemer[9], Alison J. Coffey[10], Nicolas Encina[11,12,13], Alessandra Ferlini[14], Janbernd Kirschner[15], Bianca E. Russell[16], Laurent Servais[17,18], Kristen L. Sund[19], Ryan J. Taft[10], Petros Tsipouras[20], Hana Zouk[21,22,23], ICoNS Gene List Contributors, David Bick[24], the International Consortium on Newborn Sequencing (ICoNS),
Robert C. Green[5,12,23,25]*, Nina B. Gold[26]*†

* These authors contributed equally to the manuscript.
† Corresponding author

[1]KU Leuven; [2]Boston Children's Hospital; [3]Massachusetts General Hospital; [4]Harvard Medical School; [5]Brigham and Women's Hospital; [6]Stanford School of Medicine; [7]Harvard Medical School, Department of Biomedical Informatics; [8]National Institutes of Health, National Institute of Allergy and Infectious Disease; [9]University of Liege, CHU Liege; [10]Illumina Inc.; [11]ICoNS; [12]Ariadne Labs; [13]Harvard T.H. Chan School of Public Health; [14]University of Ferrara, Department of Medical Sciences, Department of Medical Sciences, Unit of Medical Genetics; [15]University Medical Center Freiburg, Department of Neuropediatrics and Muscle Disorders; [16]University of California, Los Angeles, David Geffen School of Medicine, Department of Human Genetics, Division of Clinical Genetics; [17]University of Oxford; [18]University of Liege; [19]Nurture Genomics; [20]FirstSteps-BNSI; [21]Massachusetts General Hospital, Department of Pathology, Laboratory for Molecular Medicine; [22]Harvard Medical School, Department of Pathology; [23]Broad Institute; [24]Genomics England; [25]Mass General Brigham; [26]Massachusetts General Hospital, Department of Pediatrics; Harvard Medical School, Department of Pediatrics

Address correspondence to: Nina B. Gold, MD, Mass General Hospital for Children, Division of Medical Genetics and Metabolism, 175 Cambridge Street, Boston, MA 02114, [ngold@mgh.harvard.edu]

Structured Abstract

Purpose: Over 30 international studies are exploring newborn sequencing (NBSeq) to expand the range of genetic disorders included in newborn screening. Substantial variability in gene selection across programs exists, highlighting the need for a systematic approach to prioritize genes.

Methods: We assembled a dataset comprising 25 characteristics about each of the 4,390 genes included in 27 NBSeq programs. We used regression analysis to identify several predictors of inclusion, and developed a machine learning model to rank genes for public health consideration.

Results: Among 27 NBSeq programs, the number of genes analyzed ranged from 134 to 4,299, with only 74 (1.7%) genes included by over 80% of programs. The most significant associations with gene inclusion across programs were presence on the US Recommended Uniform Screening Panel (inclusion increase of 74.7%, CI: 71.0%-78.4%), robust evidence on the natural history (29.5%, CI: 24.6%-34.4%) and treatment efficacy (17.0%, CI: 12.3%-21.7%) of the associated genetic disease. A boosted trees machine learning model using 13 predictors achieved high accuracy in predicting gene inclusion across programs (AUC = 0.915, R^2 = 84%).

Conclusion: The machine learning model developed here provides a ranked list of genes that can adapt to emerging evidence and regional needs, enabling more consistent and informed gene selection in NBSeq initiatives.

Keywords: newborn screening; genomic sequencing; gene selection; gene-disorder associations; machine learning

Introduction

A decade ago, the BabySeq Project piloted newborn and childhood sequencing (NBSeq), a process designed to detect risk for a wide range of genetic disorders in apparently healthy infants.[1–9] Per the data repository Rx-Genes, over 700 genetic disorders now have targeted treatments or consensus guidelines for long-term management, which has further fueled enthusiasm for NBSeq.[10,11] Stakeholders, including diverse groups of parents,[12–14] rare disease specialists,[11] primary care physicians,[15] genetic counselors,[11,16,17] and the public[18,19] now support the implementation of genomic newborn screening for at least some disorders. At least 30 international research programs and companies are actively exploring this screening approach,[20–22] most of which are exchanging best practices through the International Consortium on Newborn Sequencing (ICoNS).[23]

Historically, the criteria established by Wilson and Jungner[24] have provided a framework for selecting the disorders to include in public newborn screening programs. These criteria prioritize the inclusion of childhood-onset disorders that are treatable if diagnosed in their earliest stages and require immediate intervention to prevent irreversible damage. Given the hundreds of treatable disorders that could be candidates for NBSeq and the complexity of genomic data, however, selection of the appropriate genes and disorders for NBSeq is a recurring challenge.[21,25,26] Prior studies have identified discrepancies across the genes being analyzed by a limited number of commercial NBSeq programs[27] and research studies,[28,29] but little is known about the values and variables that underlie these differences. Understanding which genes have high concordance across programs may guide emerging NBSeq research programs as they select which genes and variants to report to participants. Furthermore, the characteristics of these genes and their associated disorders can be used more empirically to prioritize candidate genes for public health programs.

To understand the variability among newborn sequencing programs, we compared the genes currently selected for analysis by 27 research studies and commercial NBSeq programs. For each gene that was included in any NBSeq program, we assembled a dataset of 25 associated characteristics. We then used a multivariate regression analysis to identify which of these characteristics were associated with inclusion across programs. Finally, we used a boosted trees model to generate a ranked list of genes, offering a data-driven approach to the prioritization of genetic disorders for population-wide NBSeq.

Methods

Study design

This cross-sectional study involved four stages: (1) identification of gene lists across international NBSeq research studies and commercial programs, (2) compilation of a dataset of characteristics for each gene included in any program, (3) statistical comparison of gene lists across programs, and (4) development of a machine learning model to predict gene inclusion across NBSeq programs based on gene characteristics.

Identification of lists of genes from research studies and commercial programs

A total of 35 NBSeq programs were identified through membership in the International Consortium on Newborn Sequencing (ICoNS) and an online search (Supplementary Table 1, Figure 1).[5,30–47] We obtained gene lists from 27 NBseq programs, including 20 research studies: BabyDetect,[30,31] BabyScreen+,[28] BabySeq,[2] BeginNGS,[32,33] Chen et al. 2023,[34] Early Check,[35,48] FirstSteps, the Generation study, gnSTAR,[36,38] GUARDIAN study,[42,49,50] Jian et al. 2022,[37] Lee et al. 2019,[43] Luo et al. 2020,[44] NeoExome,[47] NeoSeq,[40] NESTS,[41] NewbornsInSA, Progetto Genoma Puglia, Screen4Care,[45] and Wang et al. 2023.[39] Seven lists of genes from commercial firms that offer genomic newborn screening were included: FORESITE 360 (Fore Genomics), Fulgent (Newborn Genetic Analysis), Igenomix (Igenomix newborn screening), Mendelics (Teste da Bochechinha), Nurture Genomics, PerkinElmer Revvity (Genetic InsightPanel),[46] and Sema4 (Natalis).[30] Nine studies reported screening outcomes and we also summarized the positive screen rates and positive predictive values for these studies (Table 1, Supplementary Methods).

Compilation of a dataset of gene-disorder characteristics

All obtained gene lists were aggregated, standardizing gene names to HGNC nomenclature and linking each gene to a single disorder using OMIM and ClinGen (Supplementary Methods). A data repository with 25 characteristics for each gene-disorder pair was created, collecting data from five research papers and five existing databases (see Supplementary Methods).[2,10,11,32,51–53] More information on the source and description of each of the characteristics can be found in the Supplementary Methods and Supplementary Table 3. This data repository as well as the gene list data were made available online in a Github Repository.

Statistical comparison of gene lists across programs

Descriptive statistics for each gene list, including the length of the list, proportion of genes in each clinical category, and the number of genes associated with RUSP conditions were calculated (Figure 2A, 2B, Supplementary Figure 2, and Supplementary Table 4). To provide information on the concordance across all lists of genes, an UpSet plot, Venn Diagrams, and Jaccard similarity indices were provided (Figure 2C and 2D, Supplementary Figure 3 and 4). Inclusion of high concordance genes, as well as genes associated with core and secondary RUSP conditions, were plotted (Figure 2E, 2F, 2G).

A linear regression model was used to identify factors associated with inclusion in multiple gene lists. Two types of regressions were performed: regressions in which the outcome variable is the proportion of gene list inclusion *across all NBSeq programs* (Figure 3A, Supplementary Table 5 and 6) and regressions in which the outcome variable is the inclusion of a gene for each *individual* study (Figure 3B and Supplementary Table 7).

Development of a machine learning model

We developed a machine learning prediction model to prioritize genes for population-wide

NBSeq programs. Out of 25 potential gene-disorder characteristics, we selected 13 characteristics as predictors in our model: the RUSP category, clinical area, evidence base, severity of the disorder, the treatment efficacy, penetrance, treatment acceptability, age of onset, existence of an orthogonal test, the recommendation score, inheritance, prevalence, and the ClinGen Disease Validity. The remaining 12 gene and disease characteristics were excluded due to a high amount of missing data. For example, the ClinGen actionability scores were not used due to their availability for only 242 genes. Additionally, when characteristics from different sources described similar concepts, we selected the characteristic that includes data on the most genes.

Three machine learning models were compared: linear regression, random forest and boosted trees. We randomly split the gene list data into a 80% training set (n=91,312, 80% of all 114,140 potential instances of 4,390 genes included on a gene list across 26 NBSeq programs) and a 20% test set to estimate prediction power of the model. The boosted trees model, which had the highest area under the curve (AUC), was then selected to generate predictions for all 4,390 genes, providing a ranked list of genes (Supplementary Table 8).

Results

Overview of NBSeq programs

Of the 35 research and commercial programs identified, 10 were located in North America, 10 in Asia, nine in Europe, four in Australia and New Zealand, one in South America and one in Africa (Figure 1, Supplementary Table 1). The research programs anticipate a combined total sample size of 560,410 infants, with the intended enrollment in each study varying from 48 to 101,000 infants.[43,54] Nine NBSeq research programs have published or presented the screening results from a collective total of 68,628 infants by September 2024 (Table 1, Supplementary Methods).

[Figure 1 about here]

The percentage of positive screening results ranged from 1.85% in BabyDetect (3,847 infants screened for 405 genes) to 9.43% in BabySeq (159 infants screened for 4,299 genes), with an average of 3.82% positive results across 68,628 infants. There was a significant positive correlation between the percentage of positive screening results in a program and the number of genes they screened (pearson correlation coefficient of 0.653, p=0.041). A majority of the collective 1,937 positive screening results across seven studies for which detailed results were available were due to variants in *G6PD* (56.8%).

[Table 1 about here]

Four studies reported the clinical outcomes of infants who had undergone NBSeq, allowing for the calculation of these studies' positive predictive value (PPV), which varied from 12% to 79%[49,50] with an average across studies of 41% (39% when weighted by sample size).

Description of gene lists across NBSeq programs

A total of 23 programs have published or made available criteria for selecting genes and disorders for screening (Supplementary Table 2). The number of genes included in 27 programs ranged from 134 to 4,299 (median=306) (Figure 2A, 2B). A total of 4,390 genes (out of a total of 4,966 genes associated with human disease)[55] were included across at least one of the 27 gene lists (Supplementary Table 4). Of these, 4,033 genes (91.8%) were associated with a phenotype in the OMIM database. Collectively, most genes were linked to inherited metabolic disorders (IMDs) (25.4%), neurologic (15.5%), and immunologic (11.9%) disorders (Figure 2A, 2B).

[Figure 2 about here]

Discordance among gene lists used in NBSeq programs

A pairwise Jaccard Index indicated that, aside from those from commercial laboratories, most pairs of gene lists from NBSeq research programs were highly discrepant (Figure 2C and 2D, Supplementary Figure 2, 3). Of the 4,390 genes included in at least one NBSeq program, the vast majority were included by only a small number of NBSeq programs: 4,089 genes (93%) were included by 10 or fewer programs and 3,793 (87%) genes were included by five or fewer programs (Supplementary Figure 1).

Genes with high concordance across NBSeq programs

Despite this variability across gene lists, we found 74 genes (1.7% of 4,390) that were included by over 80% (22 of 27) of NBSeq programs (Supplementary Figure 1). Of these genes, 58 were associated with diseases on the US Recommended Uniform Screening Panel (RUSP) (Figure 2E, 2F). A total of 34 genes not linked to disorders on the RUSP appeared on 20 or more lists (Figure 2G).

Predictors of gene inclusion across NBSeq programs

Genes associated with core or secondary disorders on the RUSP were significantly more likely to be included in NBSeq programs (regression coefficient 74.7%, 95% confidence interval (CI): 0.710-0.784, p<0.01; regression coefficient 60.0%, 95% CI: 0.557-0.643, p<0.01, Figure 3A, Supplementary Table 5). Additionally, genes that were recommended for inclusion in newborn screening by 80% or more rare disease experts in a recent survey[11] were 43.5% (95% CI: 37.4%-49.6%, p<0.01) more likely to be included than genes that were recommended by fewer experts (Supplementary Figure 5).

[Figure 3 about here]

A strong predictor of inclusion across NBSeq programs was the ASQM evidence base, previously defined as a metric measuring a combination of gene-disease validity, published descriptions of the natural history of disease, and the availability of practice guidelines for disease diagnosis and management.[51] Genes with the highest evidence base were 29.5% more likely (95% CI: 24.6%-34.4%, p<0.01) to be included in NBSeq programs than those with less available evidence.

Other characteristics associated with inclusion across NBSeq programs were high efficacy of disease treatment (17.0%, 95% CI: 12.3%-21.7%, p<0.01), high penetrance (15.5%, 95% CI: 9.8%-21.2%, p<0.01), neonatal- or infantile-onset (15.2%, 95% CI: 10.7%-19.7%, p<0.01), high disease severity (14.5%, 95% CI: 8.2%-20.8%, p<0.01), high acceptability of treatment (with regard to the burdens and risks placed on the individual) (15.0%, 95% CI: 10.5%-19.5%, p<0.01), and the existence of an non-molecular test that could be used to confirm the diagnosis (13.8%, 95% CI: 9.3%-18.3%, p<0.01). There was variability in the importance of different characteristics across programs, but programs adhered positive value to all these characteristics with few exceptions (Figure 3B and Supplementary Table 7).

Measuring evolving knowledge about genes and diseases

We conducted a multivariate regression analysis to predict how changes in specific variables, such as treatability and evidence base, would individually influence the overall regression (Supplementary Table 6). Notably, the introduction of a new, highly acceptable treatment for a disorder with no previous treatment is associated with an increase in the likelihood of inclusion in NBSeq programs by 9.7% (95% CI: 0.1%-19.3%, p<0.05). Similarly, improving knowledge related to the natural history of a gene-disorder pair from “none” to “perfect” would increase the likelihood of inclusion in NBSeq programs by 14.8% (95% CI: 4.8%-24.8%, p<0.01).

Machine learning prediction model

Of three machine learning methods, the boosted trees model demonstrated the highest accuracy in the hold-out test set, with an area under the curve (AUC) of 0.915 and Rsquared of 80% (n=22,828, 20%) (Figure 3C, 3D). The relative importance of all variables in the boosted trees model was highest for characteristics such as the proportion of experts who recommended inclusion of the gene in NBSeq on a recent survey,[11] RUSP classification, and disease prevalence, confirming the results from the regression analysis (Supplementary Figure 8).

We used the boosted trees model to predict the observed inclusion of genes across all NBSeq programs, resulting in a list of all genes that had appeared in any NBSeq program, ranked by their predicted inclusion probabilities. Given that most genes associated with conditions on the RUSP were highly ranked, a list with these genes redacted was considered to be more illustrative of the novel capabilities of the model (Table 2). This analysis identified five genes *(PTPRC, ACADSB, NHEJ1, CYBA, and GRHPR)* that, despite being highly ranked by the model, were only included in a low proportion of NBSeq programs. To address missing data for some genes that were included across multiple NBSeq programs, we also created a second ranked list that combines the rankings generated by our machine learning model with the proportion of NBSeq programs in which each gene was observed with equal weights (Supplementary Table 8). By integrating these two sources of information, this hybrid list leverages the most comprehensive evidence available to prioritize genes for potential implementation in public health programs.

[Table 2 about here]

Discussion

NBSeq is a rapidly advancing area of global research exploring the impacts of early diagnosis for infants at risk for genetic disorders. With positive screening results in 1.85% to 9.43% of infants and a higher average PPV than some traditional newborn screening techniques,[49,56] findings from NBSeq research programs support the premise that this approach could improve early detection rates for a wide range of treatable disorders.

However, selecting the appropriate genes for screening is a critical step toward implementing population-wide NBSeq. This decision will have significant implications for at-risk infants, their families, and pediatric healthcare systems.[20,21] In this study, we compared and collected data on the genes being analyzed by 27 NBSeq programs, then developed a machine learning model to predict the inclusion of a gene across NBSeq programs. By combining this model with observed data from NBSeq programs, we generated a ranked list of genes that offers a data-driven approach to prioritizing genetic disorders for public health programs looking to incorporate NBSeq into their screening strategies.

Similar to the findings of smaller studies,[28,29] our comparison of gene lists from 27 NBSeq programs revealed substantial heterogeneity, which we probed using a series of regression models. We found that the importance of individual gene and disease characteristics varied across studies, potentially due to differences in the international prevalence of disorders, the availability of specialists and treatments in different countries and healthcare systems, or the specific goals of individual NBSeq programs. Unexpectedly, 357 genes with no disease association on OMIM and 52 with limited or refuted gene-disease validity scores from ClinGen were included by some NBSeq programs, demonstrating variation among programs in willingness to include candidate genes or those with new associations to disease.

Despite variations in the gene lists used by NBSeq programs, many share a common focus on certain clinical areas and specific genes. All programs included a substantial proportion of genes associated with disorders that are on the RUSP, reflecting the potential for genomic sequencing to detect cases missed by traditional newborn screening programs.[57–59] Of note, genes associated with some disorders on the RUSP, such as 3-methyl-crotonyl-CoA carboxylase deficiency (*MCCC1*, *MCCC2*), were widely included across lists despite not conforming to the historic Wilson-Jungner criteria, due to low penetrance and often mild symptoms.[60] This suggests that some NBSeq programs have anchored their lists around the RUSP even when the disorders are neither severe nor highly treatable.[61] Therefore, the observed concordance of a gene across NBSeq programs alone may not be sufficient to evaluate the suitability of a gene for population-wide screening.

Many gene and disease characteristics emerged as highly associated with the inclusion of genes across NBSeq programs, including the strength of published data on the natural history of disease, estimated penetrance, and the effectiveness of the associated treatment. Interestingly, despite the inclusion criteria that several NBSeq programs reported, characteristics such as the age of onset and disease severity were weakly associated with inclusion, possibly due to their subjective nature.

The machine learning model developed in this study identifies the disorders that may be most appropriate for genomic newborn screening, based on 13 characteristics and their inclusion across 27 NBSeq programs. This ranked list, along with the preferences of rare disease experts[11] could be used to prioritize genes for screening, which could then be manually curated by a team of expert reviewers. At this time, the model's predictions reflect a consensus drawn from NBSeq studies and databases, but in the future, these could be combined with hard-coded gene selection criteria. The model's flexibility also allows updates based on regional preferences, new data, or emerging therapeutics.

Importantly, this model identified several genes included by only a few NBSeq programs, but which have characteristics that are highly associated with inclusion across programs. For example, although *PTPRC*, a gene associated with severe combined immunodeficiency (SCID), was only included by 12 of 27 NBSeq programs, the model ranked it 113 of 4,390 genes. This is likely because *PTPRC* is associated with a severe immunologic disorder that typically presents in childhood and can be treated with an early hematopoietic stem cell transplant, but is a rare cause of SCID.[62] This finding highlights the model's potential to identify genes that may have been overlooked by researchers during the gene selection process.

Our study has several limitations. First, some of the NBSeq programs may have dynamic gene lists that have changed over time. Second, the treatability of various conditions varies based on country. Next, the Jaccard index may exaggerate discrepancies between gene lists of different lengths. For the regression and machine learning models, we consolidated metrics including the ASQM, BabySeq, and ClinGen databases, most of which rely on expert-curated information, such as age of onset, for which definitions vary, and data were missing among the 25 characteristics that we collected. These gaps may automatically reduce the boosted tree model's predicted inclusion of genes that are not well-characterized, resulting in lower inclusion rates for genes related to disorders that are rare or have limited published evidence. To mitigate the effects of missing data, we designed the model to be easily updated, and provided a ranked gene list that takes into account observed inclusion in addition to our model estimates. The ranked list may suffer from overfitting, given that it includes genes on which the model was trained. Lastly, the model incorporates data from funded research programs and commercial products, which may not align with the goals or constraints of public health newborn screening programs.

In the future, we plan to further develop the dataset that we have created to improve our machine learning model, by incorporating the perspectives of key informants, such as parents[13,14] and pediatricians. Additional informatics data, such as gene length, gnomAD constraint score,[63] the ENCODE blacklist,[64] or the number of associated PubMed publications, will be updated in future versions. We also plan to develop a method to iteratively add genes with newly established gene-disease validity based on recent publications to our dataset. Future research will identify genes that are not currently included in any NBSeq program, but share similar characteristics to those that are highly represented. A knowledge graph, such as PrimeKG,[65] could be used to find genes that share common molecular pathways, treatments, or symptoms to genes that are currently included in our dataset.

In summary, the growing international interest in genomic newborn screening has prompted important questions about which genes and disorders should be considered for inclusion. Due to the substantial variation in the genes included by 27 NBSeq programs, we developed an evidence-based approach to considering gene selection that draws from a comprehensive data repository encompassing over 4,000 genes. Rather than creating a static list of genes for universal implementation, our dynamic ranking system is adaptable and can be updated as new knowledge about genes, disorders, and therapeutics emerges. This work will support gene selection for both research and public health programs considering the use of population-wide NBSeq.

Data Availability

All datasets generated and/or analyzed in this study, along with the Stata, R, and Python code necessary to replicate the results, are available at https://github.com/tjminten/gim_genelist.

Funding Statement

This work was supported by the following grants: T32GM007748 (S.B.), R01HG011773 (N.G.), K08HG012811-01 (N.B.G.), TR003201 (N.B.G.), HD077671 (R.C.G.) TR003201 (R.C.G.), and EU-IMI H2020 GRANT 101034427 (A.F., J.K.).

Author Contributions

Conceptualization: L.M.A., D.B., F.B., A.J.C., N.E., A.F., N.B.G., R.C.G., J.K., T.M., B.E.R., K.L.S., L.S., R.J.T. Data curation: S.B., A.J.C., T.M., K.L.S, H.Z. Formal analysis: N.B.G., T.M. Funding acquisition: R.C.G, L.S., P.T. Investigation: S.A., S.B., N.B.G., T.M. Methodology: N.G., N.B.G., R.C.G., T.M. Resources: H.Z. Project administration: T.M. Software: T.M. Supervision: N.B.G., R.C.G., L.S. Visualization: N.G., N.B.G., T.M. Writing-original draft: S.A., N.B.G., T.M. Writing-review & editing: all authors.

Ethics Declaration

This study did not involve experiments on human participants or animals. All outcome data utilized in this research were aggregated and obtained from previously published sources, which are cited within the manuscript. As no individualized or patient-specific data were used, Institutional Review Board (IRB) or Research Ethics Committee (REC) approval were not required.

Conflict of Interest

L.M.A., A.J.C. and R.J.T. are employees and shareholders at Illumina Inc. N.G. is co-founder and equity owner of Datavisyn. N.B.G. provides occasional consulting services to RCG

Consulting and receives honoraria from Ambry Genetics. R.C.G. has received compensation for advising the following companies: Allelica, Atria, Fabric, Genomic Life and Juniper Genomics; and is co-founder of Genome Medical and Nurture Genomics. B.E.R. and K.L.S. are consultants at Nurture Genomics. L.S. received personal compensation from Zentech and Illumina Inc. P.T. is a co-founder of PlumCare RWE, LLC.

Supplemental File Listing

ICoNS Gene List Contributors authors

Mattia Gentile, Paola Orsini, Romina Ficarella, Maria Luisa Valente, Emanuela Ponzi, Athina Ververi, Maria Koutsogianni, Huang Xinwen, Xiao Rui, Zhao Zhengyan, Matthew J. Pelo, Jovanka King, Carol Siu, Karin Kassahn, Stefaan Sansen, Enrico Bertini, Aldona Zygmunt

International Consortium on Newborn Sequencing (ICoNS) authors

Sophia Adelson
Emanuele Agolini
Aljazi Al-Maraghi
Karla Alex
Fowzan Alkuraya
Ammira Alshabeeb Akil
Munira Alshehri
Derek Ansel
Niki Armstrong
Matthew Aujla
Don Bailey
Mei Baker
Jorune Balciuniene
Andrew Barry
Bruce Bennetts
Melissa Berenger
Jonathan Berg
Donna Bernstein
Arindam Bhatatcharjee
Sucheta Bhatt
David Bick
Tracey Bishop
Asaf Bitton
François Boemer
Natasha Bonhomme
George Bowley
Brenna Boyd
Heiko Brennenstuhl
Steven Brenner
Mairead Bresnahan

Mattia Gentile
Jessica Giordano
Ulrich Glumer Jensen
David Godler
Nina Gold
Aaron Goldenberg
Katie GoldenGrant
Cassie Goldman
José Manuel González de Aledo-Castillo
Daniel Gottlieb
Robert Green
Christopher Greene
Brooke Greenstein
Scott Grosse
Annette Grueters
Gulcin Gumus
Kelly Hagman
Kevin Hall
Aymeric Harmant
Sally Hartmanis
Robin Hayeems
Rose Heald
Madhuri Hegde
Rebecca Heiner-Fokkema
Lidewij Henneman
Becca Hernan
Charlotte Hobbs
Ingrid Holm
Layla Horwitz
Zhanzhi Hu

Mette Nyegaard
Justin O'Sullivan
Jelili Ojodu
Paola Orsini
Andrea Oza
Katrina Paleologos
Richard Parad
Holly Peay
Matthew Pelo
Carolyn Philstrom
Dominique Pichard
Amanda Pichini
Holly Pickering
Michelle Pirreca
Malgorzata Ponikowska
Amy Ponte
Andreas Posch
Cynthia Powell
Liana Protopsaltis
Yeyson Quevedo
Marianna Raia
Rebecca Reimers
Andy Rohrwasser
Paul Rollier
Lene Rottensten
Irakli Rtskhiladze
Nabihah Sachedina
George Sahyoun
Aditi Satija
Christian Schaaf

Thomas Brewster
PJ Brooks
Katya Broomberg
Amy Brower
Gemma Brown
James Buchanan
Caleb Bupp
Candance Cameron
Lauren Capacchione
Diana Carli
Onassis Castillo Ceballo
Kee Chan
Jillian Chance
Georgia Charalambidou
Winnie Chen
Yun-Ru Chen
Wendy Chung
Brian Chung
Megan Clarke
Susan Clasper
F. Sessions Cole
Heidi Cope
Stephanie Coury
Tony Cox
Tamara Dangouloff
Earnest James Paul Daniel
Katrin Eivindardottir Danielsen
Emeline Davoine
Tom Defay
Geethanjali Devadoss Gandhi
Joseph Dewulf
Lisa Diller
Pakhi Dixit
Martijn Dolle
Lilian Downie
Erin Drake
Suzanne Drury
Annelotte Duintjer
Bugrahan Duz
Maria Iascone
Ken Irvine
Guanjun Jin
Kelsey Kalbfleisch
Ines Kander
Lucy Kaplun
Dalia Kasperaviciute
Karin Kassahn
Leni Kauko
Riina Kaukonen
Nicole Kelly
Dhayo Khangsar
Jovanka King
Clare Kingsley
Stephen Kingsmore
Brian Kirmse
Rachel Klein
Stefan Koelker
Youssef Kousa
Elizaveta Krupoderova
Paul Kruszka
Katherine Langley
Ciara Leckie
Emmanuelle Lecommandeur
David Ledbetter
Pamela Lee
Beomhee Lee
Camille Level
Celine Lewis
Anna Lewis
Ruby Liu
Mauro Longoni
Alberte Lundquist
Sebastian Lunke
Kate MacDuffie
Ankit Malhotra
Lionel Marcelis
Maria Martinez-Fresno
Gert Matthijs
Jennifer Schleit
Richard Scott
Lauren Scully
Stacey Seeloff
Laurent Servais
Nidhi Shah
Maija Siitonen
Sikha Singh
Carol Siu
Hadley Smith
Lisa Sniderman King
Neal Sondheimer
Lourdes St George
Zornitza Stark
Robert Steiner
Ulrik Stoltze
Asbjørg Stray-Pedersen
Kristen Sund
Paris Tafas
Polakit Teekakirikul
Dimitrios Thanos
Audrey Thurm
Meekai To
Petros Tsipouras
Alice Tuff-Lacey
Heather Turner
Philip Twiss
Fiona Ulph
Daniel Uribe
Tiina Urv
Cora Vacher
Kris Van Den Bogaert
Mirjam van der Burg
Eva Van Steijvoort
Yiota Veloudi
Elizabeth Vengoechea
Els Voorhoeve
Martin Vu
Melissa Wasserstein

David Eckstein
Matthew Ellinwood
Katarzyna Ellsworth
Sarah Elsea
Nicolas Encina
Harriet Etheredge
Laurence Faivre
Alessandra Ferlini
Monica Ferrie
Terri Finkel
Petra Furu
Jamie Galarza-Cornejo
Ya Gao
Judit Garcia-Villoria
Liz Gardner
Amy Gaviglio
Michael Gelb

Roberts Melbardis
Jessica Merritt
Radja Messai Badji
Lou Metherell
Nanna Balle Mikkelsen
Laura Milko
Nicole Miller
Thomas Minten
Sian Morgan
Katarzyna Mosiewicz
Ulrike Mütze
Sukhvinder Nicklen
Minna Niemela
Dau-Ming Niu
Sarah Norris
Antonio Novelli
Arwa Nusair

Michael Watson
Bryn Webb
Anna Wedell
Thomas Westover
Alexandra Wiedemann
Meredith Wright
Cindy Wu
Julie Yeo
Nancy Yin-Hsiu Chien
Shamila Yusuff
Tomasz Zemojtel
Bethany Zettler
Zhengyan Zhao
Joanna Ziff
Rebekah Zimmerman
Michela Zuccolo

References

1. Waisbren SE, Bäck DK, Liu C, et al. Parents are interested in newborn genomic testing during the early postpartum period. *Genet Med*. 2015;17(6):501-504.

2. Ceyhan-Birsoy O, Machini K, Lebo MS, et al. A curated gene list for reporting results of newborn genomic sequencing. *Genet Med*. 2017;19(7):809-818.

3. Genetti CA, Schwartz TS, Robinson JO, et al. Parental interest in genomic sequencing of newborns: enrollment experience from the BabySeq Project. *Genet Med*. 2019;21(3):622-630.

4. Holm IA, Agrawal PB, Ceyhan-Birsoy O, et al. The BabySeq project: implementing genomic sequencing in newborns. *BMC Pediatr*. 2018;18(1):225.

5. Ceyhan-Birsoy O, Murry JB, Machini K, et al. Interpretation of Genomic Sequencing Results in Healthy and Ill Newborns: Results from the BabySeq Project. *Am J Hum Genet*. 2019;104(1):76-93.

6. Pereira S, Smith HS, Frankel LA, et al. Psychosocial Effect of Newborn Genomic Sequencing on Families in the BabySeq Project: A Randomized Clinical Trial. *JAMA Pediatr*. 2021;175(11):1132-1141.

7. Pereira S, Robinson JO, Gutierrez AM, et al. Perceived Benefits, Risks, and Utility of Newborn Genomic Sequencing in the BabySeq Project. *Pediatrics*. 2019;143(Suppl 1):S6-S13.

8. Green RC, Shah N, Genetti CA, et al. Actionability of unanticipated monogenic disease risks in newborn genomic screening: Findings from the BabySeq Project. *Am J Hum Genet*. 2023;110(7):1034-1045.

9. Smith HS, Zettler B, Genetti CA, Hickingbotham MR, Coleman TF, Lebo M, Nagy A, Zouk H, Mahanta L, Christensen KD, Pereira S, Shah ND, Gold NB, Walmsley S, Edwards S, Homayouni R, Krasan GP, Hakonarson H, Horowitz CR, Gelb BD, Korf BR, McGuire AL, Holm IA, Green RC. The BabySeq Project: A Clinical Trial of Genome Sequencing in a Diverse Cohort of Infants. *Am J Hum Genet, In Press*. Published online 2024.

10. Bick D, Bick SL, Dimmock DP, Fowler TA, Caulfield MJ, Scott RH. An online compendium of treatable genetic disorders. *Am J Med Genet C Semin Med Genet*. 2021;187(1):48-54.

11. Gold NB, Adelson SM, Shah N, et al. Perspectives of Rare Disease Experts on Newborn Genome Sequencing. *JAMA Netw Open*. 2023;6(5):e2312231.

12. Joseph G, Chen F, Harris-Wai J, Puck JM, Young C, Koenig BA. Parental Views on Expanded Newborn Screening Using Whole-Genome Sequencing. *Pediatrics*. 2016;137 Suppl 1(Suppl 1):S36-S46.

13. Timmins GT, Wynn J, Saami AM, Espinal A, Chung WK. Diverse Parental Perspectives of the Social and Educational Needs for Expanding Newborn Screening through Genomic Sequencing. *Public Health Genomics*. Published online September 15, 2022:1-8.

14. Gold NB, Omorodion JO, Del Rosario MC, et al. Preferences of parents from diverse

backgrounds on genomic screening of apparently healthy newborns. *J Genet Couns*. Published online October 28, 2024. doi:10.1002/jgc4.1994

15. Acharya K, Ackerman PD, Ross LF. Pediatricians' attitudes toward expanding newborn screening. *Pediatrics*. 2005;116(4):e476-e484.

16. Cao M, Notini L, Ayres S, Vears DF. Australian healthcare professionals' perspectives on the ethical and practical issues associated with genomic newborn screening. *J Genet Couns*. 2023;32(2):376-386.

17. del Rosario MC, Swenson KB, Coury S, Schwab J, Green RC, Gold NB. Genetic counselors' perspectives on genomic screening of apparently healthy newborns in the United States. *Genetics in Medicine Open*. 2024;(101885):101885.

18. Bombard Y, Miller FA, Hayeems RZ, et al. Public views on participating in newborn screening using genome sequencing. *Eur J Hum Genet*. 2014;22(11):1248-1254.

19. Lynch F, Best S, Gaff C, et al. Australian Public Perspectives on Genomic Newborn Screening: Risks, Benefits, and Preferences for Implementation. *Screening*. 2024;10(1). doi:10.3390/ijns10010006

20. Stark Z, Scott RH. Genomic newborn screening for rare diseases. *Nat Rev Genet*. 2023;24(11):755-766.

21. Downie L, Halliday J, Lewis S, Amor DJ. Principles of Genomic Newborn Screening Programs: A Systematic Review. *JAMA Netw Open*. 2021;4(7):e2114336.

22. Bros-Facer V, Taylor S, Patch C. Next-generation sequencing-based newborn screening initiatives in Europe: an overview. *Rare Dis Orphan Drugs J*. 2023;2:21.

23. International Consortium on Newborn Sequencing. https://www.iconseq.org/

24. Wilson JMG, Jungner G. *The Principles and Practice of Screening for Disease*.; 1966.

25. Bick D, Ahmed A, Deen D, et al. Newborn Screening by Genomic Sequencing: Opportunities and Challenges. *Screening*. 2022;8(3). doi:10.3390/ijns8030040

26. Baple EL, Scott RH, Banka S, et al. Exploring the benefits, harms and costs of genomic newborn screening for rare diseases. *Nat Med*. 2024;30(7):1823-1825.

27. DeCristo DM, Milko LV, O'Daniel JM, et al. Actionability of commercial laboratory sequencing panels for newborn screening and the importance of transparency for parental decision-making. *Genome Med*. 2021;13(1):50.

28. Downie L, Bouffler SE, Amor DJ, et al. Gene selection for genomic newborn screening: moving towards consensus? *Genet Med*. Published online January 23, 2024:101077.

29. Betzler IR, Hempel M, Mütze U, et al. Comparative analysis of gene and disease selection in genomic newborn screening studies. *J Inherit Metab Dis*. Published online May 16, 2024. doi:10.1002/jimd.12750

30. Dangouloff T, Hovhannesyan K, Piazzon F, et al. Baby detect: Universal genomic newborn screening for early, treatable, and severe conditions. *J Neurol Sci*. 2023;455. doi:10.1016/j.jns.2023.121259

31. Website. "Population-Based, First-Tier Genomic Newborn Screening in a Single Maternity Ward in Belgium: Results of Babydetect Project." n.d. Accessed August 8,

2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4896054

32. Kingsmore SF, Smith LD, Kunard CM, et al. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. *Am J Hum Genet*. 2022;109(9):1605-1619.

33. Owen MJ, Lefebvre S, Hansen C, et al. An automated 13.5 hour system for scalable diagnosis and acute management guidance for genetic diseases. *Nat Commun*. 2022;13(1):4057.

34. Chen T, Fan C, Huang Y, et al. Genomic Sequencing as a First-Tier Screening Test and Outcomes of Newborn Screening. *JAMA Netw Open*. 2023;6(9):e2331162.

35. Bailey DB Jr, Gehtland LM, Lewis MA, et al. Early Check: translational science at the intersection of public health and newborn screening. *BMC Pediatr*. 2019;19(1):238.

36. Huang X, Wu D, Zhu L, et al. Application of a next-generation sequencing (NGS) panel in newborn screening efficiently identifies inborn disorders of neonates. *Orphanet J Rare Dis*. 2022;17(1):66.

37. Jian M, Wang X, Sui Y, et al. A pilot study of assessing whole genome sequencing in newborn screening in unselected children in China. *Clin Transl Med*. 2022;12(6):e843.

38. Yang RL, Qian GL, Wu DW, et al. A multicenter prospective study of next-generation sequencing-based newborn screening for monogenic genetic diseases in China. *World J Pediatr*. 2023;19(7):663-673.

39. Wang X, Sun Y, Guan XW, et al. Newborn genetic screening is highly effective for high-risk infants: A single-centre study in China. *J Glob Health*. 2023;13:04128.

40. Wang H, Yang Y, Zhou L, Wang Y, Long W, Yu B. NeoSeq: a new method of genomic sequencing for newborn screening. *Orphanet J Rare Dis*. 2021;16(1):481.

41. Hao C, Guo R, Hu X, et al. Newborn screening with targeted sequencing: a multicenter investigation and a pilot clinical study in China. *J Genet Genomics*. 2022;49(1):13-19.

42. Chung WK, Kanne SM, Hu Z. An Opportunity to Fill a Gap for Newborn Screening of Neurodevelopmental Disorders. *Screening*. 2024;10(2). doi:10.3390/ijns10020033

43. Lee H, Lim J, Shin JE, et al. Implementation of a Targeted Next-Generation Sequencing Panel for Constitutional Newborn Screening in High-Risk Neonates. *Yonsei Med J*. 2019;60(11):1061-1066.

44. Luo X, Sun Y, Xu F, et al. A pilot study of expanded newborn screening for 573 genes related to severe inherited disorders in China: results from 1,127 newborns. *Ann Transl Med*. 2020;8(17):1058.

45. Ferlini A, Gross ES, Garnier N, Screen4Care consortium. Rare diseases' genetic newborn screening as the gateway to future genomic medicine: the Screen4Care EU-IMI project. *Orphanet J Rare Dis*. 2023;18(1):310.

46. Balciuniene J, Liu R, Bean L, et al. At-Risk Genomic Findings for Pediatric-Onset Disorders From Genome Sequencing vs Medically Actionable Gene Panel in Proactive Screening of Newborns and Children. *JAMA Netw Open*. 2023;6(7):e2326445.

47. Cao Z, He X, Wang D, et al. Targeted exome sequencing strategy (NeoEXOME) for Chinese newborns using a pilot study with 3423 neonates. *Mol Genet Genomic Med*.

2024;12(1):e2357.

48. Cope HL, Milko LV, Jalazo ER, et al. A systematic framework for selecting gene-condition pairs for inclusion in newborn sequencing panels: Early Check implementation. *Genet Med*. Published online October 4, 2024:101290.

49. Chung W, Ziegler A, Koval-Burt C, et al. O35: Feasibility of expanded newborn screening using genome sequencing for early actionable conditions in a diverse city. *Genetics in Medicine Open*. 2024;2(101369):101369.

50. Ziegler A, Koval-Burt C, Kay D, et al. Expanded newborn screening using genome sequencing for early actionable conditions. *JAMA*. Published online October 24, 2024. doi:10.1001/jama.2024.19662

51. Milko LV, O'Daniel JM, DeCristo DM, et al. An Age-Based Framework for Evaluating Genome-Scale Sequencing Results in Newborn Screening. *J Pediatr*. 2019;209:68-76.

52. Berg JS, Foreman AKM, O'Daniel JM, et al. A semiquantitative metric for evaluating clinical actionability of incidental or secondary findings from genome-scale sequencing. *Genet Med*. 2016;18(5):467-475.

53. Rehm HL, Berg JS, Brooks LD, et al. ClinGen--the Clinical Genome Resource. *N Engl J Med*. 2015;372(23):2235-2242.

54. Spiekerkoetter U, Bick D, Scott R, et al. Genomic newborn screening: Are we entering a new era of screening? *J Inherit Metab Dis*. 2023;46(5):778-795.

55. Amberger JS, Bocchini CA, Schiettecatte FJM, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D789-98.

56. Kwon C, Farrell PM. The magnitude and challenge of false-positive newborn screening test results. *Arch Pediatr Adolesc Med*. 2000;154(7):714-718.

57. Wojcik MH, Zhang T, Ceyhan-Birsoy O, et al. Discordant results between conventional newborn screening and genomic sequencing in the BabySeq Project. *Genet Med*. 2021;23(7):1372-1375.

58. Adhikari AN, Gallagher RC, Wang Y, et al. The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nat Med*. 2020;26(9):1392-1397.

59. Cook S, Dunn E, Kornish J, et al. Molecular testing in newborn screening: VUS burden among true positives and secondary reproductive limitations via expanded carrier screening panels. *Genet Med*. 2023;26(4):101055.

60. Arnold GL, Koeberl DD, Matern D, et al. A Delphi-based consensus clinical practice protocol for the diagnosis and management of 3-methylcrotonyl CoA carboxylase deficiency. *Mol Genet Metab*. 2008;93(4):363-370.

61. Forsyth R, Vockley CW, Edick MJ, et al. Outcomes of cases with 3-methylcrotonyl-CoA carboxylase (3-MCC) deficiency - Report from the Inborn Errors of Metabolism Information System. *Mol Genet Metab*. 2016;118(1):15-20.

62. Roberts JL, Buckley RH, Luo B, et al. CD45-deficient severe combined immunodeficiency caused by uniparental disomy. *Proc Natl Acad Sci U S A*. 2012;109(26):10456-10461.

63. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

64. Amemiya HM, Kundaje A, Boyle A. The ENCODE blacklist: Identification of problematic regions of the genome. *Sci Rep*. 2019;9. doi:10.1038/s41598-019-45839-z

65. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. *Sci Data*. 2023;10(1):67.

Figure Legends

Figure 1. Research and commercial genomic newborn screening (NBSeq) programs. Gene lists from 27 of these programs were included in the analysis (denoted with an asterisk). Intended enrollment sizes are indicated where available.

Figure 2. Description, concordance and content of gene lists of genomic newborn screening programs.

A. Clinical areas of 4,299 genes included in BabySeq.
B. Counts and clinical areas of genes included in 26 research and commercial genomic newborn screening programs (n=4,390).
C. Jaccard similarity index, which offers a quantitative comparison of how closely related the gene lists are.
D. UpSet plot of gene lists of 4 large research studies. The matrix below the bar graph represents each individual study and their intersections (n=818).
E. Inclusion of genes associated with core Recommended Uniform Screening Panel (RUSP) conditions. The x-axis is each genomic newborn screening program and y-axis are individual genes; the corresponding cell is colored if the gene is included on a given list.
F. Inclusion of genes associated with secondary RUSP conditions.
G. Inclusion of genes on 20 lists or more that are not associated with RUSP conditions.

Figure 3. Determinants and prediction model of gene inclusion in genomic newborn screening. In a and b, RUSP category (n=4,474), survey recommendation and orthogonal test (n=649), evidence base, efficacy, penetrance, disease severity, treatment acceptability and neonatal or infant onset (n=749). ROC, Receiver Operating Characteristic; AUC, Area Under the Curve.

A. Regression coefficients (and confidence intervals) associated with various gene and disease characteristics predicting inclusion across gene lists.
B. Heat map with regression coefficients associated with gene and disease characteristics for each individual genomic newborn screening program.
C. ROC curves for three prediction models in the hold-out test set (n=895 genes).
D. Scatter plot of predicted versus observed gene list inclusion, showing the fit of the boosted trees model on the 20% hold-out set (n=895 genes).

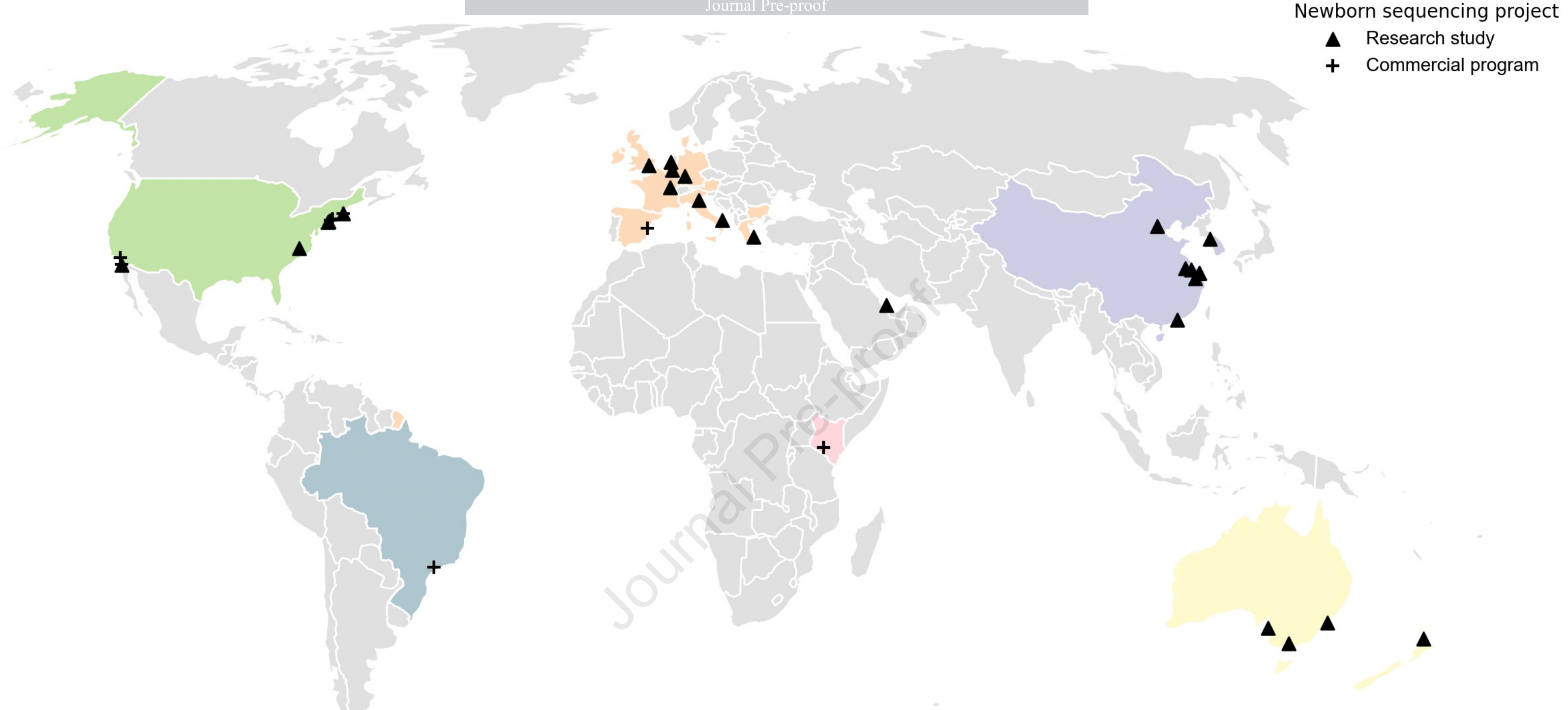
Table 1. Percentage of positive results from genomic newborn screening research programs. Positive cases of *G6PD*, as well as follow-up data for positively screened infants reported where available. PPV, positive predictive value.

		Sequencing data				Follow-up data		
NBSeq research study	Genes	Infants sequenced	G6PD deficiency	Total positive	Total positive (in %)	Total follow-up	Total diagnosed	PPV
BabyDetect	405	3,847	44	71	1.85%			
PerkinElmer Panel	268	606	4	13	2.15%			
Wang et al. 2023	164	10,334		232	2.25%	231	50	22%
gnSTAR	134	4,986	40	113	2.27%			
Chen et al. 2023	142	29,601	689	813	2.75%	797	402	50%
GUARDIAN Study	237	4,000		147	3.68%	151	120	79%
NESTS	465	11,484	338	902	7.85%	414	50	12%
PerkinElmer GS	6,000	562	4	46	8.19%			
NeoExome	601	3,049		271	8.89%			
BabySeq	4,299	159	1	15	9.43%			
TOTAL		**68,628**		**2,623**	**3.82%**			

Table 2. List of 50 genes with highest predicted inclusion across NBSeq programs, excluding genes on the Recommend Uniform Screening Panel (RUSP).
ASQM, Age-Based Semi Quantitative Metric.

Rank	Gene	Disorder	OMIM Phenotype	Clinical area	BabySeq category	ClinGen validity	ClinGen Actionability (/12)	ASQM score (/15)	Expert Recommendation	Observed list inclusion (/27)	Predicted list inclusion
13	*G6PC1*	Glycogen storage disease Ia	232200	Metabolism	A			13	93%	26	96%
17	*OTC*	Ornithine transcarbamylase deficiency	311250	Metabolism	A	Definitive	10AB	14	98%	26	96%
40	*SLC37A4*	Glycogen storage disease Ib	232220	Metabolism	A			13	93%	25	91%
41	*CPS1*	Carbamoylphosphate synthetase I deficiency	237300	Metabolism	A	Definitive	9AB	13	86%	25	90%
47	*SLC25A15*	Hyperornithinemia-hyperammonemia-homocitrullinemia syndrome	238970	Metabolism	A	Definitive		10	85%	25	88%
51	*AGL*	Glycogen storage disease IIIa	232400	Metabolism	A	Definitive		12	85%	24	86%
52	*NAGS*	N-acetylglutamate synthase deficiency	237310	Metabolism	A	Definitive	10NC	14	85%	22	86%
55	*ALPL*	Hypophosphatasia, infantile	241500	Endocrinology	A	Definitive	10CC	11	71%	23	84%
56	*GALNS*	Mucopolysaccharidosis IVA	253000	Metabolism	A	Definitive	9CA	9	87%	23	84%
57	*ALDOB*	Fructose intolerance, hereditary	229600	Metabolism	A		10NC	14	84%	23	83%
60	*BTK*	Agammaglobulinemia, X-linked 1	300755	Immunology	A	Definitive			81%	24	82%
65	*F9*	Hemophilia B	306900	Hematology	A	Definitive	9CB	14	90%	23	81%
68	*CYBB*	Chronic granulomatous disease, X-linked	306400	Immunology	A			12	74%	25	80%
69	*G6PD*	Hemolytic anemia, G6PD deficient (favism)	300908	Hematology	A	Definitive	7DC	12	80%	23	80%
73	*CYP11B1*	Adrenal hyperplasia, congenital, due to 11-beta-hydroxylase deficiency	202010	Endocrinology	A			11	92%	21	79%
74	*SMPD1*	Niemann-Pick disease	257200	Metabolism	A	Definitive		9	85%	21	78%
75	*ARSB*	Mucopolysaccharidosis type VI (Maroteaux-Lamy)	253200	Metabolism	A	Definitive		11	92%	22	77%
76	*CTNS*	Cystinosis, nephropathic	219800	Nephrology	A	Definitive	9CB	14	80%	20	77%
77	*ATP7B*	Wilson disease	277900	Metabolism	A	Definitive	11CA	12	81%	21	77%
78	*CYBA*	Chronic granulomatous disease 4, autosomal recessive	233690	Immunology	A			12	74%	18	77%
79	*ABCC8*	Hyperinsulinemic hypoglycemia, familial, 1	256450	Endocrinology	A	Definitive		14	81%	20	75%
80	*RB1*	Retinoblastoma	180200	Oncology	A	Definitive	10CB	13	89%	21	75%
81	*ALDH7A1*	Epilepsy, early-onset, 4, vitamin B6-dependent	266100	Neurology		Definitive	11CB	12	86%	21	74%
82	*TG*	Thyroid dyshormonogenesis 3	274700	Endocrinology	A			14		18	73%
84	*TPO*	Thyroid dyshormonogenesis 2A	274500	Endocrinology	A			14		18	71%
85	*PHEX*	Hypophosphatemic rickets, X-linked dominant	307800	Endocrinology			9CC		74%	20	71%
86	*TH*	Segawa syndrome, recessive	605407	Neurology	A	Definitive		13	70%	20	71%
87	*ATP7A*	Menkes disease	309400	Metabolism	A	Definitive	9CA	12	84%	19	71%
89	*SLC2A1*	GLUT1 deficiency syndrome 1, infantile onset, severe	606777	Metabolism	A	Definitive		11	90%	19	71%
90	*GLA*	Fabry disease	301500	Metabolism	A	Definitive	9CA	12	83%	20	70%
91	*DUOX2*	Thyroid dyshormonogenesis 6	607200	Endocrinology	A			13		20	70%
92	*WAS*	Wiskott-Aldrich syndrome	301000	Immunology	A	Definitive	9CC		68%	20	69%
93	*FBP1*	Fructose-1,6-bisphosphatase deficiency	229700	Metabolism				13	80%	18	69%
94	*PYGL*	Glycogen storage disease VI	232700	Metabolism	A			10	83%	19	69%
96	*SPR*	Dystonia, dopa-responsive, due to sepiapterin reductase deficiency	612716	Neurology	A	Definitive		12	70%	20	69%
97	*CD40LG*	Immunodeficiency, X-linked, with hyper-IgM	308230	Immunology	A	Definitive		11	74%	17	68%
98	*LIPA*	Wolman disease	620151	Metabolism	A	Definitive		7	80%	19	67%
100	*GUSB*	Mucopolysaccharidosis VII	253220	Metabolism	A	Definitive		9	89%	19	67%
101	*SCNN1B*	Pseudohypoaldosteronism, type IB2, autosomal recessive	620125	Endocrinology	A			14	49%	17	66%
102	*POU1F1*	Pituitary hormone deficiency, combined or isolated, 1	613038	Endocrinology	A			12	66%	18	66%
103	*TSHR*	Hypothyroidism, congenital, nongoitrous, 1	275200	Endocrinology	A			12		24	66%
104	*AVPR2*	Diabetes insipidus, nephrogenic, 1	304800	Endocrinology	A			11	60%	18	66%
105	*KCNJ11*	Diabetes mellitus, transient neonatal 3	610582	Endocrinology	A	Definitive		12	76%	18	64%
106	*GATM*	Cerebral creatine deficiency syndrome 3	612718	Metabolism		Definitive		10	85%	17	64%

108	*GLUD1*	Hyperinsulinism-hyperammonemia syndrome	606762	Metabolism	A	Definitive	11	87%	18	63%
110	*PHKB*	Phosphorylase kinase deficiency of liver and muscle, autosomal recessive	261750	Metabolism	A	Definitive	13	76%	17	63%
111	*CYP17A1*	17,20-lyase deficiency, isolated	202110	Endocrinology				90%	18	62%
114	*NPC1*	Niemann-Pick disease	257220	Metabolism	A	Definitive	12	69%	19	62%
115	*SLC7A7*	Lysinuric protein intolerance	222700	Metabolism	A	Definitive	12	85%	16	62%
116	*PROP1*	Pituitary hormone deficiency, combined, 2	262600	Endocrinology	A		13	66%	19	62%



North America

- ▲ BabySeq* (MA), 100,000
- ▲ BeginNGS* (CA), 2,000
- ▲ Early Check* (NC), 10,000
- ▲ GUARDIAN* (NY), 100,000
- ▲ ScreenPlus (NY)
- + FORESITE 360* (CA)
- + Fulgent* (CA)
- + Nurture Genomics* (MA)
- + PerkinElmer* (CT)
- + Sema4* (CT)

Europe

- ▲ BabyDetect* (BEL), 40,000
- ▲ FirstSteps* (GRE), 101,000
- ▲ Generation* (ENG), 100,000
- ▲ Screen4Care* (EU), 18,000
- ▲ Genoma Puglia* (ITA), 6,000
- ▲ Perigenomed (FRA), 20,000
- ▲ Cradle (NED)
- ▲ New_Lives (GER)
- + Igenomix* (ESP)

Asia

- ▲ Chen et al.* (CHI), 29,601
- ▲ gnSTAR* (CHI), 4,986
- ▲ Jian et al.* (CHI), 321
- ▲ Lee et al.* (KOR), 48
- ▲ Luo et al.* (CHI), 1,127
- ▲ NeoExome* (CHI), 3,423
- ▲ NeoSeq* (CHI), 196
- ▲ NESTS* (CHI), 11,484
- ▲ Wang et al.* (CHI), 10,224
- ▲ Sidra Medicine (QAT)

South America

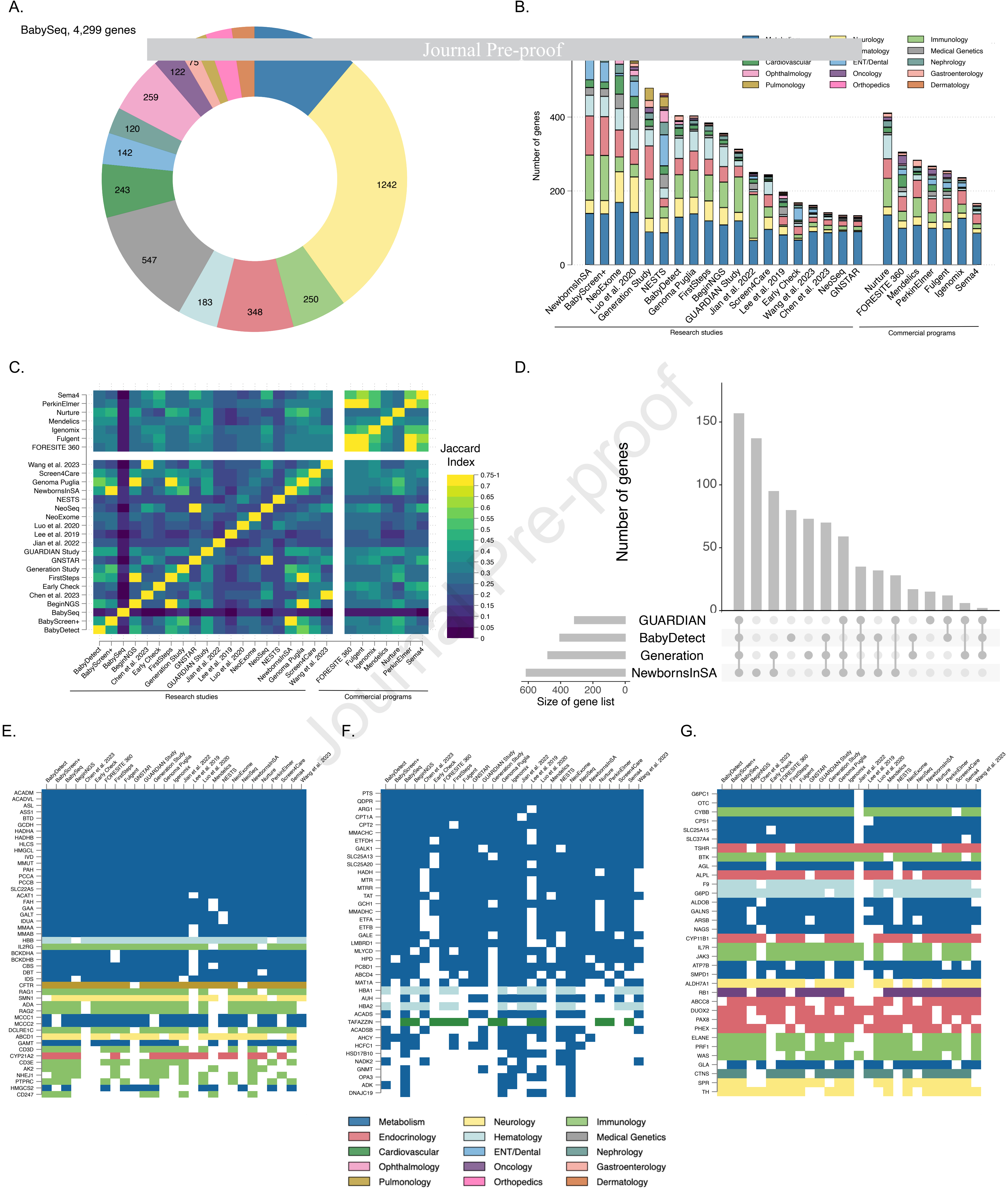
- + Mendelics* (BRA)

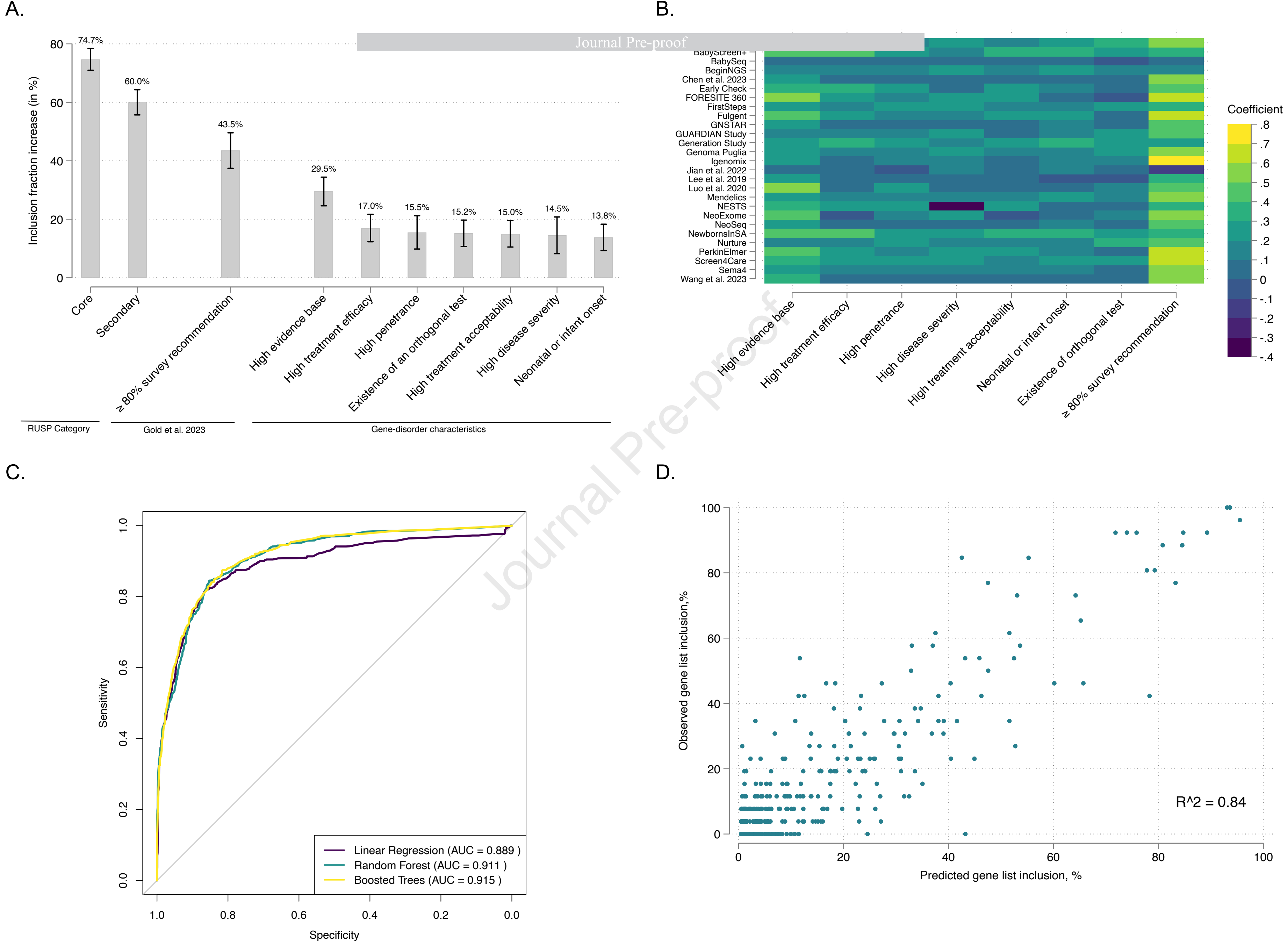
Africa

- + KIBs (KEN)

Oceania

- ▲ BabyScreen+* (AUS), 1,000
- ▲ NewbornsInSA* (AUS), 1,000
- ▲ Newborn Gen Prog (NZ)
- ▲ TRAIL (AUS)


A.
BabySeq, 4,299 genes
1242
547
348
250
183
243
142
120
259
122
75
B.
Number of genes
0
200
400
NewbornsInSA
BabyScreen+
NeoExome
Luo et al. 2020
Generation Study
NESTS
BabyDetect
Genoma Puglia
FirstSteps
BeginNGS
GUARDIAN Study
Jian et al. 2022
Screen4Care
Lee et al. 2019
Early Check
Wang et al. 2023
Chen et al. 2023
NeoSeq
GNSTAR
Nurture
FORESITE 360
Mendelics
PerkinElmer
Fulgent
Igenomix
Sema4
Research studies
Commercial programs
Metabolism
Neurology
Immunology
Hematology
Medical Genetics
Cardiovascular
ENT/Dental
Nephrology
Ophthalmology
Oncology
Gastroenterology
Pulmonology
Orthopedics
Dermatology
C.
Sema4
PerkinElmer
Nurture
Mendelics
Igenomix
Fulgent
FORESITE 360
Wang et al. 2023
Screen4Care
Genoma Puglia
NewbornsInSA
NESTS
NeoSeq
NeoExome
Luo et al. 2020
Lee et al. 2019
Jian et al. 2022
GUARDIAN Study
GNSTAR
Generation Study
FirstSteps
Early Check
Chen et al. 2023
BeginNGS
BabySeq
BabyScreen+
BabyDetect
Jaccard Index
0.75-1
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Research studies
Commercial programs
D.
Number of genes
0
50
100
150
GUARDIAN
BabyDetect
Generation
NewbornsInSA
600
400
200
0
Size of gene list
E.
ACADM
ACADVL
ASL
ASS1
BTD
GCDH
HADHA
HADHB
HLCS
HMGCL
IVD
MMUT
PAH
PCCA
PCCB
SLC22A5
ACAT1
FAH
GAA
GALT
IDUA
MMAA
MMAB
HBB
IL2RG
BCKDHA
BCKDHB
CBS
DBT
IDS
CFTR
RAG1
SMN1
ADA
RAG2
MCCC1
MCCC2
DCLRE1C
ABCD1
GAMT
CD3D
CYP21A2
CD3E
AK2
NHEJ1
PTPRC
HMGCS2
CD247
F.
PTS
QDPR
ARG1
CPT1A
CPT2
MMACHC
ETFDH
GALK1
SLC25A13
SLC25A20
HADH
MTR
MTRR
TAT
GCH1
MMADHC
ETFA
ETFB
GALE
LMBRD1
MLYCD
HPD
PCBD1
ABCD4
MAT1A
HBA1
AUH
HBA2
ACADS
TAFAZZIN
ACADSB
AHCY
HCFC1
HSD17B10
NADK2
GNMT
OPA3
ADK
DNAJC19
G.
G6PC1
OTC
CYBB
CPS1
SLC25A15
SLC37A4
TSHR
BTK
AGL
ALPL
F9
G6PD
ALDOB
GALNS
ARSB
NAGS
CYP11B1
IL7R
JAK3
ATP7B
SMPD1
ALDH7A1
RB1
ABCC8
DUOX2
PAX8
PHEX
ELANE
PRF1
WAS
GLA
CTNS
SPR
TH
Metabolism
Neurology
Immunology
Endocrinology
Hematology
Medical Genetics
Cardiovascular
ENT/Dental
Nephrology
Ophthalmology
Oncology
Gastroenterology
Pulmonology
Orthopedics
Dermatology


A.
Inclusion fraction increase (in %)
80
60
40
20
0
74.7%
60.0%
43.5%
29.5%
17.0%
15.5%
15.2%
15.0%
14.5%
13.8%
Core
Secondary
≥ 80% survey recommendation
High evidence base
High treatment efficacy
High penetrance
Existence of an orthogonal test
High treatment acceptability
High disease severity
Neonatal or infant onset
RUSP Category
Gold et al. 2023
Gene-disorder characteristics
B.
BabyScreen+
BabySeq
BeginNGS
Chen et al. 2023
Early Check
FORESITE 360
FirstSteps
Fulgent
GNSTAR
GUARDIAN Study
Generation Study
Genoma Puglia
Igenomix
Jian et al. 2022
Lee et al. 2019
Luo et al. 2020
Mendelics
NESTS
NeoExome
NeoSeq
NewbornsInSA
Nurture
PerkinElmer
Screen4Care
Sema4
Wang et al. 2023
High evidence base
High treatment efficacy
High penetrance
High disease severity
High treatment acceptability
Neonatal or infant onset
Existence of orthogonal test
≥ 80% survey recommendation
Coefficient
.8
.7
.6
.5
.4
.3
.2
.1
0
-.1
-.2
-.3
-.4
C.
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Linear Regression (AUC = 0.889)
Random Forest (AUC = 0.911)
Boosted Trees (AUC = 0.915)
D.
Observed gene list inclusion,%
0
20
40
60
80
100
Predicted gene list inclusion, %
0
20
40
60
80
100
R^2 = 0.84