npj | Science of Learning

www.nature.com/npjscilearn

ARTICLE OPEN

Genetic associations with mathematics tracking and persistence in secondary school

K. Paige Harden[1,8*], Benjamin W. Domingue[2,8], Daniel W. Belsky[3], Jason D. Boardman[4], Robert Crosnoe[5], Margherita Malanchini[1], Michel Nivard[6], Elliot M. Tucker-Drob[1] and Kathleen Mullan Harris[7]

Maximizing the flow of students through the science, technology, engineering, and math (STEM) pipeline is important to promoting human capital development and reducing economic inequality. A critical juncture in the STEM pipeline is the highly cumulative sequence of secondary school math courses. Students from disadvantaged schools are less likely to complete advanced math courses. Here, we conduct an analysis of how the math pipeline differs across schools using student *polygenic scores*, which are DNA-based indicators of propensity to succeed in education. We integrated genetic and official school transcript data from over 3000 European-ancestry students from U.S. high schools. We used polygenic scores as a molecular tracer to understand how the flow of students through the high school math pipeline differs in socioeconomically advantaged versus disadvantaged schools. Students with higher education polygenic scores were tracked to more advanced math already at the beginning of high school and persisted in math for more years. Analyses using genetics as a molecular tracer revealed that the dynamics of the math pipeline differed by school advantage. Compared to disadvantaged schools, advantaged schools buffered students with low polygenic scores from dropping out of math. Across all schools, even students with exceptional polygenic scores (top 2%) were unlikely to take the most advanced math classes, suggesting substantial room for improvement in the development of potential STEM talent. These results link new molecular genetic discoveries to a common target of educational-policy reforms.

npj Science of Learning (2020)5:1 ; https://doi.org/10.1038/s41539-020-0060-2

INTRODUCTION

Math matters for economic success.[1] American students who take math courses beyond Algebra 2 are more likely to enroll in college and complete a STEM degree[2–4] and have better labor market outcomes.[5–7] Students from low-income families and schools are less likely to take advanced math courses in secondary school, which impairs their entry to post-secondary STEM education and ultimately to a STEM career.[8–10] There are, however, continuing debates about whether the underrepresentation of low-income students in STEM is due to the diminished resources available to their schools and families or, rather, due to those students having lower aptitude or interest in math.[8,11–14] Despite the intense focus on STEM outcomes, it is challenging to conduct rigorous studies of whether and how schools differ in the flow of students through the math pipeline. In particular, analyses that statistically control for traditional measures of student aptitude or interest might lead to biased conclusions about how the math pipeline differs across schools, because these student characteristics can themselves be influenced by previous educational experiences.[8]

Our project addresses the challenge of understanding how students' progress through the STEM pipeline might vary as a function of school characteristics by using a DNA-based measure of students' likelihood to succeed in education. A previous genome-wide association study (GWAS) of 1.1 million people identified hundreds of genetic variants associated with higher educational attainment.[15] These results can be used to calculate an *education polygenic score* (education-PGS), which is a composite index of genetic variants associated with completing more years of school.[16–18] The education-PGS predicts whether or not an individual completes college about as well as his/her family income does.[15] Moreover, unlike traditional measures of student aptitude, individual differences in genetic sequence are fixed at conception and cannot be changed by educational experiences.

Polygenic scores can therefore be used as a molecular tracer to measure flows of students through the STEM pipeline and assess how these flows differ across schools. Just as a radiologist might administer a radioactive tracer to track the flow of blood within the body, researchers can use genetics as a molecular tracer to get a clearer image of how students progress through the twists and turns of the educational system. Here, we use polygenic scores to follow the curricular histories of students who attended secondary schools with varying levels of socioeconomic advantage. This approach offers a way of diagnosing the extent to which students who have high genetic propensities for success in education leak out of the STEM pipeline by failing to advance in their mathematics training.

In mapping the flow of students through the secondary school math curriculum, we focus on two dimensions of high school mathematics coursetaking—*tracking* and *persistence*. In some countries (e.g., Germany), students are tracked into different types of secondary schools at a discrete number of branch points. The U.S., in contrast, does not have a formal tracking system. Instead, students are offered curricular options that are differentiated by content and difficulty (e.g., Pre-Algebra vs. Algebra I vs. Algebra II). Students are informally *tracked* toward final math credentials via their course placement in the first year of secondary school (or

[1]Department of Psychology and Population Research Center, University of Texas at Austin, Austin, TX, USA. [2]Graduate School of Education, Stanford University, Stanford, CA, USA. [3]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA. [4]Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CA, USA. [5]Department of Sociology and Population Research Center, University of Texas at Austin, Austin, TX, USA. [6]Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands. [7]Department of Sociology and Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [8]These authors contributed equally: K. Paige Harden, Benjamin W. Domingue. *email: harden@utexas.edu

earlier).[19] As subsequent coursetaking hinges on successful completion of pre-requisites and mastery of cumulative content, students' curricular decisions become strongly path-dependent.[19–21] Students additionally vary after the first year in whether they *persist* in their track throughout secondary school, move to a less-advanced track, or discontinue mathematics training entirely.

We first sought to validate the education-PGS as a molecular tracer by testing whether it predicted being tracked into a more advanced math class at the beginning of high school and whether it predicted persisting in math for longer. Next, we used the education-PGS to examine differences between schools in the flow of students through the STEM pipeline. Specifically, we focused on the difference between schools that served mainly students from well-educated families versus schools that served mainly students from families with less formal education. We found that the education-PGS predicts movement through the math coursetaking pipeline but that it also intersects with school characteristics. We are able to use our genetic analysis to make an important observation about the role of schools independent of genetic variation. In particular, we observe that two students with the *same* education-PGS might differ substantially in their progress through the STEM pipeline, depending on their school characteristics.

While our approach leverages insights from large-scale genetic studies, we emphasize that genetics are clearly not the only factor that matters for student achievement. Indeed, by comparing students who are equivalent in their measured genetic propensities, but who attend different schools, our analyses can illuminate how mathematics achievement depends on contextual factors. In this way, we view research that uses genetic tools as complementing educational research on how school-based factors, such as instructional practices, can boost mathematics achievement.[22]

RESULTS

Mathematics coursetaking can be categorized from school transcripts

Analyses used genetic and official school transcript data on $N =$ 3,635 unrelated adolescents from the National Longitudinal Study of Adolescent to Adult Health (Add Health, see Methods; Supplementary Fig. 1).[23] Respondents were enrolled in a U.S. high school in 1994–1995. We restricted analyses to European-ancestry participants to prevent inadvertently conflating genetic variation with racial or ethnic background. Previous analyses of national population patterns have revealed a fairly standardized sequence of math coursework, ranging from more basic courses like Pre-Algebra to more advanced courses like Calculus.[24,25] We used this sequence to categorize each participant's math coursework across four years of secondary school, based on information obtained from schools, including course catalogs, school information forms, and interviews with school administrators (Supplementary Table 2).

At the beginning of secondary school (9th grade, age ~14 years), most students were enrolled in Algebra 1 (51%), but some students were tracked to less advanced (Pre-Algebra or below, 29%) or more advanced (Geometry or above, 20%) courses. A student's final level of mathematics training was strongly dependent on 9th-grade course enrollment: 44% of those enrolled in Geometry or higher in 9th-grade ultimately completed Calculus, compared to only 4.2% of those enrolled in Algebra 1 and 1% enrolled in Pre-Algebra or lower level math class.

Student polygenic score predicts mathematics tracking

Students with higher polygenic scores were more likely to be tracked into more advanced math courses in 9th grade (Fig. 1a,

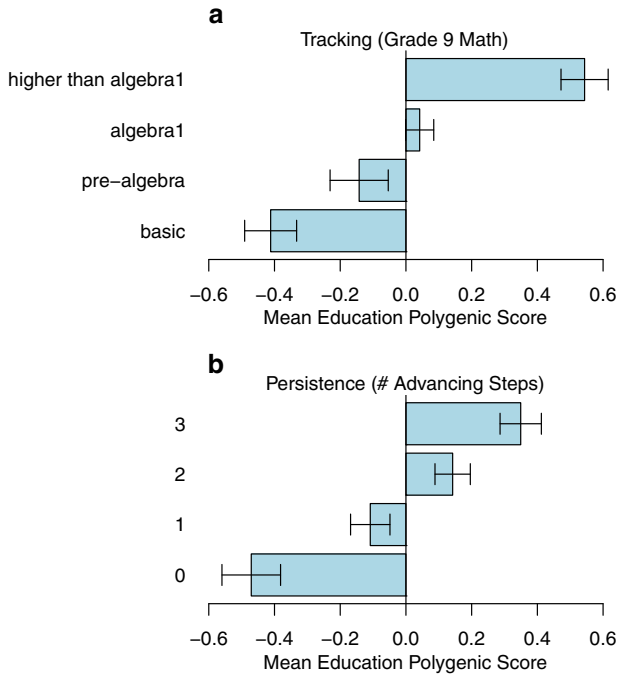

Fig. 1 Students with higher education-associated polygenic scores are tracked to more advanced math and persist for longer in math. Error bars represent 95% confidence intervals around the mean.

$b = 0.583$, SE $= 0.035$, 95% CI $= [0.516, 0.656]$, Supplementary Table 3). However, Add Health participants with a higher education-PGS more often grew up in high-SES families and attended high-SES schools, as compared to participants with lower polygenic scores.[26,27] These gene-environment correlations raise the possibility that genetic associations with mathematics tracking could be due to clustering of students with higher polygenic scores in environmental contexts that better support math achievement. To address this possibility, we repeated our analysis of tracking in the 9th grade using measures of school-SES and family SES as covariates (Methods, Supplementary Table 3). As expected, students from higher SES families were tracked to more advanced math courses at the beginning of secondary school ($b = 0.419$, SE $= 0.039$, 95% CI $= [0.344, 0.5]$). The association between school-SES and tracking was also positive but not significantly different than zero ($b = 0.704$, SE $= 0.571$, 95% CI $= [-0.671, 1.74]$). However, including family SES and school-SES as covariates attenuated the association between the education-PGS and mathematics tracking in the 9th-grade only by about 20% (attenuated from $b = 0.583$, SE $= 0.035$, to $b = 0.469$, SE $= 0.035$, 95% CI $= [0.397, 0.53]$, Supplementary Table 3). Note that the association with genetics was roughly comparable in magnitude to the association with family SES. As a stronger test of whether the genetic association with mathematics tracking was due to clustering of students with high education-PGS into certain schools, we repeated our analysis of 9th-grade tracking yet again, this time using school-fixed-effects regression to compare students to their schoolmates (Supplementary Table 3).[28] Comparing only students who were in Algebra 1 or below, students with higher education-PGS were less likely, compared to their schoolmates, to be placed in a remedial track (Pre-Algebra or lower) than in Algebra 1 ($b = 0.387$, SE $= 0.054$, 95% CI $= [0.294, 0.504]$). Similarly, comparing only students who were in Algebra 1 or above, students with higher education-PGS were more likely, compared to their schoolmates, to be placed in an advanced track

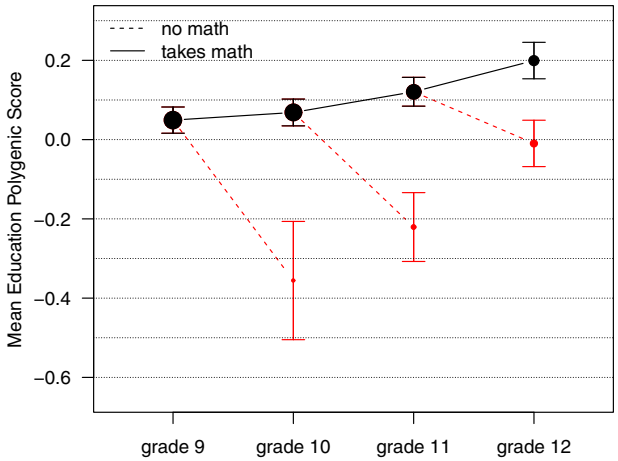

Fig. 2 Genetic associations with persistence in math recur year-after-year. Error bars represent 95% confidence intervals around the mean. Size of the dots represents number of students enrolled or not enrolled in math in each year.

(Geometry or higher) rather than in Algebra 1 ($b = 0.587$, SE = 0.047, 95% CI = [0.501, 0.681]).

Student polygenic score predicts persistence in mathematics coursetaking

What happens to students after the 9th-grade? Participants in this sample attended high school in the mid-1990 s, when the average high school graduation requirement in U.S. states was 2.4 years of math coursework.[29] Rates of math drop-out accelerated in later years of secondary school (9th-grade: 2.6%, 10th-grade: 5.2%, 11th-grade: 17.6%, 12th-grade: 44.7%; Supplementary Table 2). Once students dropped out of math, they tended to remain out of math coursework; only 8% of students enrolled in a math class after a year of no math. We summarized persistence across the four years of transcript follow-up as number of advancing steps in the math coursework sequence, ranging from zero to three. For example, a student who completed Algebra 1, Geometry, and Algebra 2 in the first three years of secondary school but who did not take a math course in the 12th-grade, took two advancing steps.

Students with higher education-PGS took more advancing steps (Fig. 1b; $b = 0.139$, SE = 0.011, 95% CI = [0.117, 0.159]; Supplementary Table 4). We then repeated this analysis using a number of additional covariates. As we observed for tracking, students from higher SES families were more likely to persist in math coursetaking across secondary school ($b = 0.120$, SE = 0.015, 95% CI = [0.089, 0.147]). The association with school-SES was similarly positive but not statistically significant ($b = 0.234$, SE = 0.0171, 95% CI = [−0.098, 0.572]). But the education-PGS association with persistence was only modestly attenuated after accounting for family- and school-SES covariates ($b = 0.096$, SE = 0.013, 95% CI = [0.072, 0.12]). We also considered a model including 9th-grade course placement as a covariate. Students who were tracked to Pre-Algebra or lower in the 9th-grade persisted less in math than those in Algebra 1 ($b = -0.221$, SE = 0.038, 95% CI = [−0.293, −0.147]). In contrast, students in more advanced math tracks in 9th-grade (Geometry or higher) did not differ from those placed in Algebra 1 ($b = -0.059$, SE = 0.044, 95% CI = [−0.147, 0.020]). Controlling for tracking in the 9th-grade, the education-PGS again remained associated with persistence ($b = 0.087$, SE = 0.012, 95% CI = [0.061, 0.11]). While these analyses suggest a robust association between the educational attainment PGS and persistence, within-family analyses suggested that the polygenic score is not predictive of the sibling with a higher PGS being more persistent than the other (see SI) although this finding could partially be a function of that sample being of limited size.

As with tracking, we repeated this analysis yet again using school fixed-effects to compare students to others in their school. Consistent with previous analyses, participants with higher education-PGSs took more advancing steps in their mathematics coursetaking than their schoolmates ($b = 0.117$, SE = 0.013, 95% CI = [0.091, 0.138]; Supplementary Table 4).

We next analyzed persistence on a year-by-year basis. As shown in Fig. 2, most students were enrolled in math in 9th-grade, and their mean education-PGS was the sample mean (i.e., zero). Few of these students dropped out of math in 10th-grade, but these early drop-outs had a low average education-PGS (less than 0.3 SD below the mean). The pace of attrition increased in subsequent years (note growth in size of the red dots), and students who continued to take any math class were an increasingly positively selected group.

We considered whether the education-PGS provided any novel information above and beyond what could be observed from students' performance in math class. It did. This set of analyses focused on students who were enrolled in any math class in the 9th-, 10th-, and 11th-grades, and tested enrollment in any math class in the subsequent year. End-of-year grade point averages (GPAs; on a 4-point scale) in math were obtained from the school transcripts. At every year, students from higher SES families, students attending higher SES schools, and students who had higher math GPAs were more likely to enroll in math the subsequent year (Supplementary Table 5). After controlling for these covariates, a 1-SD increase in the education-PGS was still associated with 1.26 times greater odds of taking a math class in 10th-grade (95% CI = [1.05–1.56]), 1.15 times greater odds in 11th-grade (95% CI = [1.08–1.28]), and 1.13 times greater odds in the 12th-grade (95% CI = [1.05–1.22]).

Our analyses reveal genetically stratified flows of students through the mathematics training pipeline. We visualized these flows using a "river plot" (Fig. 3).[30] In the river plot, participants' math courses (rows) are plotted by year of secondary education (columns). Courses are ordered from most advanced at the top of the graph to least-advanced at the bottom. The widths of the rivers (i.e., the edges connecting row-column nodes) indicate the number of students moving from one course to another. The color of the rivers represents the average education-PGS for students following a particular path (higher in blue, lower in orange). Collectively, these results support the premise that the education-PGS can be used as a molecular tracer to evaluate how students flow through the STEM pipeline in secondary school.

Higher SES schools buffer the risks faced by students predicted to struggle in math

Building on recent evidence,[26,31] we next conducted two analyses of how STEM pipeline dynamics varied by school advantage. First, we tested if the genetic association with tracking differed between high- and low-SES schools using cumulative link models with product terms to capture interactions between school-SES and the education PGS (Fig. 4a). The interaction term was positive, suggesting that the education-PGS predicted 9th-grade tracking more strongly among students in higher-status schools than in lower-status schools, but this effect was not statistically significant (interaction $b = 0.59$, SE = 0.291, 95% CI [−0.007, 1.11]); Supplementary Table 3). A student with an education-PGS of + 1 (top 16th percentile) who is in a high-status school has a 33.1% probability of being tracked to Geometry in the 9th-grade (note horizontal gray line in Fig. 4a). In order to have the same probability of being placed in Geometry, a student in a low-status school would need to have an education-PGS of + 2.0 (top 2%). Robustness analyses using non-parametric LOESS and adjacent-category logit models suggested similar patterns (see Supplementary Fig. 2).

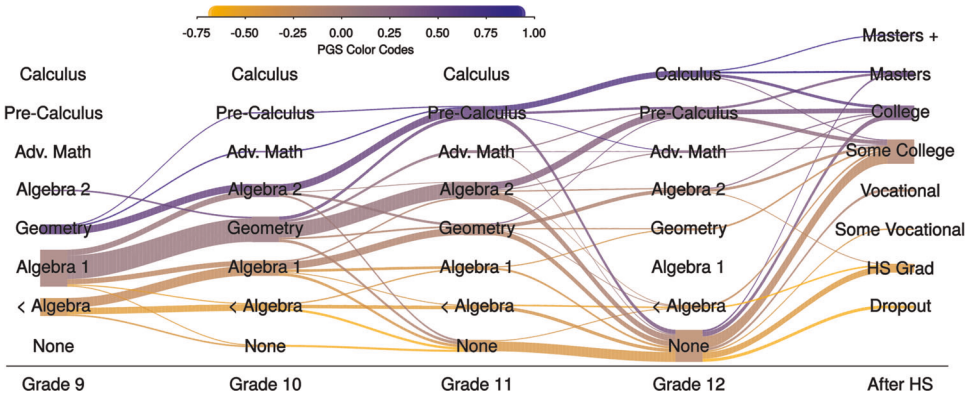

Fig. 3 Student DNA can be used to visualize the flow of students through the high school math curriculum. Columns represent year of secondary school; rows represent mathematics course sequence ranging from least to most advanced. Width of the rivers connecting columns proportional to number of students. Shading of rivers represents the average education polygenic score for students in a particular course in a particular year, ranging from low (orange) to high (blue).

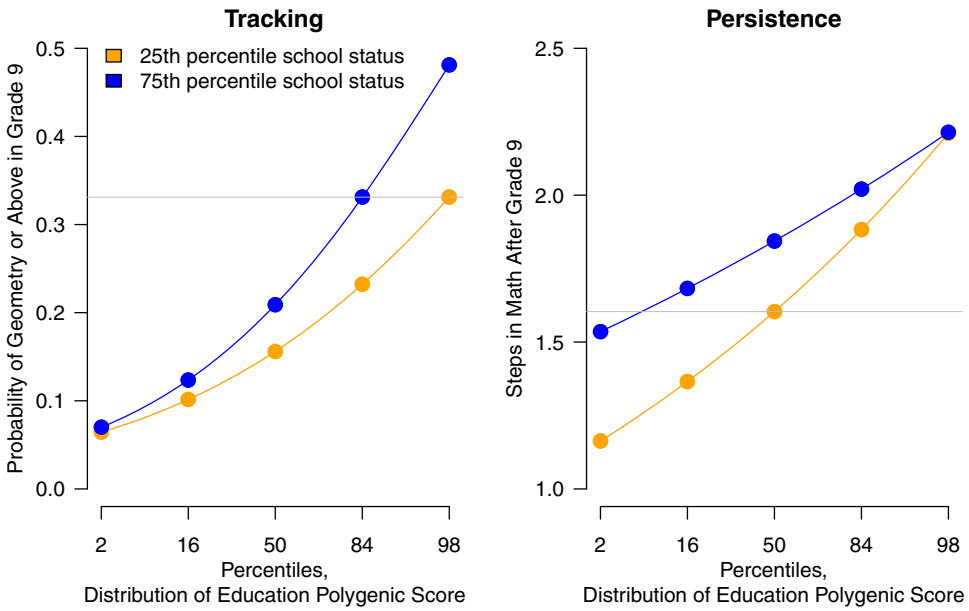

Fig. 4 Student tracking and persistence vary as a function of school. a Students with high education-associated polygenic scores are more likely to be tracked into advanced math in advantaged schools than in disadvantaged schools. Fitted probabilities of being tracked to Geometry or higher in the 9th grade, based on cumulative link logit model. School status measured by percent of students whose mother graduated from high school. See Supplementary Table 3 for full model results. **b** Students with low education-associated polygenic scores persist more in math in advantaged schools than in disadvantaged schools. Model-implied number of advancing steps from 9–to–12th-grade, based on Poisson model. At least 1 year of math beyond the 9th grade was compulsory in most U.S. states. See Supplementary Table 4 for full model results.

Second, we tested the interaction between education-PGS and school-SES in predicting number of advancing steps in math. There was a significant and negative interaction on mathematics persistence, such that low-PGS students were *less* likely to drop out of math if attending high-status schools as compared to low-status schools ($b = -0.304$, SE = 0.074, 95% CI = [−0.443, −0.147]; Supplementary Table 4). The interaction effect was similar when including 9th-grade tracking as a covariate ($b = -0.282$, SE = 0.072, 95% CI = [−0.41, −0.147]). Figure 4b shows how the number of advancing steps varied as a function of education-PGS in schools at the 0.25 quantile versus 0.75 quantile of school status. High-PGS students persisted about equally in their mathematics training regardless of school status. In contrast, students with an average or low education-PGS were particularly likely to drop out of math in low-SES schools. For example, students attending low-SES schools with an average education-PGS completed 1.6 advancing steps (note gray horizontal line in Fig. 4b). However, in high-SES schools, a similarly low level of mathematics persistence is only seen in students at the very low end of the genetic distribution (education-PGS = −1.5, bottom 7%-ile).

Our final analyses focused on school differences in whether or not a student completed Calculus, the most advanced course category in the 9-course sequence. Results from a logistic regression found that school-SES and the education-PGS each predicted taking Calculus, but they did not significantly interact (Supplementary Table 6). Students with an average education-PGS had nearly twice the chances of taking Calculus in a high-SES school (11%) than in a low-SES one (6%). Calculus was rare even among students with exceptional polygenic scores (top 2%, or +2 SD above the mean): High-PGS students had a 24% probability

of taking Calculus in a low-SES school and a 31% probability in a high-SES school.

DISCUSSION

We used data on student genetics as a molecular tracer to test how the flow of students through the high school math curriculum varied between disadvantaged versus advantaged school contexts. Students with higher education polygenic scores tended to enroll in more advanced mathematics tracks in the 9th-grade, and they were more likely to persist in these tracks through the end of high school. Furthermore, genetic associations with tracking and persistence were not explained by differences between schools or measured differences in family socioeconomic status (SES).

However, these student-level results might be complicated by the existence of gene-environment interactions, as the flow of students through the mathematics pipeline differed between high-SES and low-SES schools. Students with low polygenic scores were buffered from dropping out of math in high-status schools. Consequently, students in high-status schools had substantially better math credentials by the end of high school, compared to students who had comparable polygenic scores but who were enrolled in low-status schools. This study cannot identify specific causal factors for school-level differences. Such factors could include school resources and instructional practices (e.g.,[22]), as well as correlated features of neighborhoods and other environmental contexts. Nevertheless, whatever the cause of these school-level differences, our results underscore that many students are not going as far in mathematics as we might expect were they at another school. Furthermore, our results contrast with previous suggestions that school differences in academic outcomes might solely reflect differences in the genetic composition of their student bodies.[13]

Our findings suggest that genetics may provide a novel approach to studying challenging educational problems. A persistent methodological problem in educational research is how to separate out the effects of teachers and schools from the effects of student characteristics that are non-randomly distributed across schools, such as family income.[32] Observable student characteristics are in flux during these crucial years of development and are also associated with both upstream and downstream developmental influences, such as previous schooling. Genetics, as a fixed characteristic of the student that is as predictive of success in schooling as family income,[15] offers researchers an additional tool for studying how student development varies by context.

With the caveat that the Add Health data represents an earlier cohort of students, our results further suggest that even advantaged school contexts do a poor job of maximizing human capital. Out of students who both had exceptional polygenic scores (+2SD above the Add Health sample mean or the top 2% of the distribution) and attended high status high school, about 31% took Calculus by the end of high school, whereas only 24% of students who had the same score and attended low-status schools took calculus. This deficit in advanced coursetaking among students with exceptional genetic propensities for succeeding in education suggests that many students who likely could excel in mathematics are not being put in opportunities to do so. In terms of our running metaphor: the pipeline is leaking. Badly.

We acknowledge several limitations. First, these analyses do not inform our understanding of disparities between ethnic and racial groups, which is one of the most pressing problems in the U.S. educational system.[33] Polygenic scores are useful only for understanding individual differences between people who share the same genetic ancestry, and the validity of education GWAS results has been established only for people of European ancestry.[15] The extent to which results will generalize to other populations is uncertain. Just as the biomedical use of polygenic scores developed in European-ancestry populations has the potential to exacerbate pre-existing health disparities, using polygenic scores in educational contexts also has the potential to exacerbate pre-existing achievement gaps between racial and ethnic groups.[34]

Second, the genetic predictor deployed here captures only a fraction of the genotypic variation relevant for education. Associations with the polygenic scores are attenuated by measurement error,[35] and other variants (e.g., rare variants[36]) relevant for ultimate educational attainment might not operate through the processes described here. We anticipate that genetic associations with mathematics coursetaking and other academic achievement outcomes will increase in magnitude as the sample sizes of GWASs continue to increase and as polygenic scores incorporate information from whole-genome sequencing.[37]

Third, our analyses lacked information on specific features of policy or programming that might explain why students with similar genetics fare differently in different schools. Our results suggest that the relevant factors are linked in some way to the socioeconomic characteristics of the students served. Future research is needed to identify what modifiable features of students' educational environments are involved in amplifying or suppressing genetic influences on their skill development as they move through the STEM pipeline. We encourage researchers to investigate the panoply of institutional, social, demographic, and cultural differences that exist between schools and that might contribute to school-level differences in the association between student genetics and academic achievement.

Fourth, previous studies have suggested that up to half of the polygenic score association with educational outcomes might operate indirectly. In addition to giving information about his/her own biology, a student's polygenic score also reflects the genotype of his/her parents, which is in turn associated with the environment that the parents provide.[38] Consequently, the association between genotype and math curricular choices might partially operate not through the genetically-influenced characteristics of the student herself, but through the genetically-influenced characteristics of her parents, such as the greater knowledge that college-educated parents have about how to navigate a differentiated curriculum.[39] Work is already beginning to document such pathways.[40–42] Disentangling such indirect genetic effects[43] from genetic effects that operate through the biology of the student will require larger samples of genetic relatives, such as parent-offspring trios.[44] We conducted an initial exploration of this question by comparing siblings raised in the same family (Supplement), but the relatively small number of sibling pairs with transcript data available in Add Health limits the definitiveness of our conclusions about the role of indirect genetic effects.

Fifth, there is limited information on the educational histories of students prior to the 9th-grade, but students' secondary school experiences are, of course, shaped by their previous mathematics skill development and curricular choices. We suspect that the genetic associations with tracking in the 9th-grade partially reflects genetic variation in math skills that have been acquired prior to high school; however, the roles of attributes other than math ability, including the constellation of personality and motivational factors referred to as "non-cognitive" skills, are also likely important.[45] We see potential hints of this here, as polygenic scores predict persistence in math *even* after controlling for the student's math grades in the previous year. Other genetically influenced traits that are potentially influential for course placement are the student's attention problems, behavioral and mental health difficulties, academic interests, motivations, and self-concept, and ability to elicit support and investment from adults.[46–48]

It is now well established that educational attainment is heritable[49] and can be predicted from an individual's DNA.[50–52] What is less well-understood is *how* genetic differences between individuals lead to differences in educational outcomes. In order for genetics research to be of greater relevant to education research—and, ultimately, to education policy and practice—greater knowledge is needed about the developmental and social processes that connect students' DNA to their educational outcomes.[53] As sample sizes for GWAS continue to increase, more and more specific genetic variants associated with complex human phenotypes, like educational attainment, will continue to be identified. There are dangers associated with genetic research being used to justify an overly reductionistic or bio-deterministic account of educational outcomes.[54,55] Here, however, we illustrate how DNA measures offer new opportunities for educational science. Specifically, we show that genetics can be used to identify leaks in the STEM pipeline and can refine our understanding who is benefitting (and who is not) from educational contexts.

METHODS

Sample

The National Longitudinal Study of Adolescent to Adult Health (Add Health)[23] is a nationally-representative cohort drawn from a probability sample of 80 U.S. high schools and 52 U.S. middle schools (in roughly 90 U.S. communities). Participating schools were representative of all U.S. schools in 1994–95 with respect to region, urban setting, school size, school type, and race or ethnic background. Add Health participants provided written informed consent for participation in all aspects of Add Health in accordance with the University of North Carolina School of Public Health Institutional Review Board guidelines that are based on the Code of Federal Regulations on the Protection of Human Subjects 45CFR46. This project was approved by the Stanford University IRB (eProtocol #: 35363).

We constructed our analytic sample as follows (see also Supplementary Table 1). At Wave 1, data was collected for $N = 20{,}369$ respondents. At Wave 3, respondents of the Add Health study, who were then 18–26 years old, were contacted and asked to give signed consent for the release of their official high school transcripts (AHAA).[25] Transcripts were collected regardless of whether the student graduated from high school. At Wave 4, biospecimens were collected for genome-wide genotyping (described in[27,56]). Of those in the genetic sample, we focused on unrelated respondents of European ancestry, due to the problem of population stratification in diverse samples.[33,57] Transcripts ($N = 12{,}032$) and genetic samples ($N = 5{,}045$ of European origin) were collected for partially overlapping subsets of the Wave 1 respondents. Our analytic sample therefore consisted of 3,635 European ancestry respondents with both genotypic and transcript data.

Descriptive statistics are contained in Supplementary Table 1. Compared to the full Add Health sample, our analytic sample had higher family SES, higher overall GPAs, and higher rates of post-secondary education. Missing information further reduced sample size in some analyses.

Polygenic scores

Using results from the most recent educational attainment GWAS,[15] we constructed a polygenic score using all single nucleotide polymorphisms (SNPs) with reported effect sizes that are also in the Add Health genetic dataset and where the effect allele can be reliably matched to the allele reported in the Add Health genetic data. We residualized the PGS on the top 10 principal components of genetic ancestry and then standardized the PGS based on the full set of respondents in the genetic dataset ($N = 5045$). A similar PGS has been used in previous work.[26,27] Genotyped respondents who were not in the transcript dataset had a mean education-PGS of −0.11, whereas the genotyped respondents with transcript data had a mean education-PGS of 0.04 (Supplementary Table 1).

Transcript coursetaking and course grades

Course content information obtained from the schools was used to identify the level of each course on a student's transcript and to assign it a Classification of Secondary School Courses code. These codes were used to develop an ordinal indicator of the math course sequence, ranging from 0 (no math) to 9 (calculus). These indicators were developed by the AHAA project[25,58] to be compatible with the 2000 National Assessment of Educational Progress High School Transcript Study[24] and are based on population patterns of coursetaking as derived from the National Education Longitudinal Study of 1988.[59] The percentages of students enrolled in each level at each year are in Supplementary Table 2. For analysis of 9th-grade coursetaking, math courses that focus on remedial skills (Basic/Remedial and General) were collapsed, as were math classes beyond Geometry (Algebra 2, Pre-Calculus, Advanced Math, and Calculus).

Students' final math course grades at each year were obtained from transcripts and coded on a 0–4 scale ($0 = F$, $1 = D$, $2 = C$, $3 = B$, $4 = A$). If a student took the class pass/fail, withdrew, or received an incomplete, then his/her course grade is missing.[25] A cumulative GPA was also computed from these transcript-based grades.

Family and school SES

Family-of-origin SES was indexed using the first principal component of parental education, job status, income, and the number of benefits received (loadings were 0.58, 0.43, 0.49, and 0.49, respectively); see[27] for additional information on this indicator. The family SES variable was standardized with respect to the full Wave I sample; the current analytic sample was more advantaged than the full sample ($M = 0.34$, $SD = 1.16$).

We used an indicator of school status used previously.[26,60] Add Health administered an in-school survey to all students in participating high schools ($N = 90{,}118$). This information was used to calculate the percentage of students at each school who report that their mother graduated high school.

Analyses

Our statistical models varied as a function of the outcome variable. For non-categorical outcome variables (e.g., number of advancing steps), we fit baseline generalized linear models of the form

$$E\left(y_{ij}\right) = g^{-1}\left(b_0 + b_1\mathrm{PGS}_{ij} + \mathrm{controls}\right), \quad (1)$$

where i indexes school, j indexes individual, and an appropriate link function g is chosen given the distribution of the outcome y. For analyses of 9th-grade tracking, we fit ordered logistic regressions.[61] For analysis of persistence, we used Poisson regressions. We fit interaction models of the form:

$$E\left(y_{ij}\right) = g^{-1}\left(b_0 + b_1\mathrm{PGS}_{ij} + b_2\mathrm{School\ Status}_{ij} + \mathrm{b}_3 PGS_{\mathrm{ij}} \cdot \mathrm{School\ Status}_{ij} + \mathrm{controls}\right). \quad (2)$$

For interaction models, we also included interactions between covariates and the key main effects, so as to guard against spurious findings from specification error.[62] Thus, in models examining interactions between the education-PGS and School Status (as in Fig. 4), we also included interactions between PGS and sex, School Status and sex, PGS and birth year, and School Status and birth year (see Supplementary Table 3). Standard errors for key models (Tables S3–S6) are adjusted for school-based clustering using bootstrap resampling of the schools.[63]

For our ordinal categorical outcomes (e.g., tracking in 9th-grade), we consider cumulative link models.[61] As used here, cumulative link models assert that:

$$\Pr\left(y_{ijk} \le j\right) = f\left(\theta_j - \left(b_0 + b_1\mathrm{PGS}_{ij} + \mathrm{controls}\right)\right), \quad (3)$$

where k in [0,1,…, K] now indexes the category of the outcome y. We used a logit link, rendering this model equivalent to the proportional odds model.[64] We again compute cluster-robust standard errors using school-based bootstrap resampling. One key assumption of this model is that the effect of the predictors does not vary across categories. We therefore also present results from alternative models (e.g., adjacent-category logit models) as robustness checks, see Supplement.

Reporting summary

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

DATA AVAILABILITY

Data for this study come from the National Longitudinal Study of Adolescent to Adult Health (Add Health) and the Adolescent Health and Academic Achievement study (AHAA), which provided transcript data for Add Health participants. Restricted-use

data files with phenotypic information from Add Health and AHAA can be obtained from the UNC Carolina Population Center following the procedures outlined in http://www.cpc.unc.edu/projects/addhealth/data/restricteduse. Genetic data on Add Health participants can be obtained from dbGaP at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001367.v1.p1.

CODE AVAILABILITY

Code for analyses and figures is available in a GitHub repository at https://github.com/ben-domingue/transcripts-eaPGS.

Received: 3 June 2019; Accepted: 9 January 2020;
Published online: 05 February 2020

REFERENCES

1. Board, N. S. *Revisiting the STEM workforce: A Companion to Science and Engineering Indicators 2014* (National Science Foundation VA, 2015).
2. Aughinbaugh, A. The effects of high school math curriculum on college attendance: Evidence from the NLSY97. *Econ. Educ. Rev.* **31**, 861–870 (2012).
3. Long, M. C., Conger, D. & Iatarola, P. Effects of high school course-taking on secondary and postsecondary success. *Am. Educ. Res. J.* **49**, 285–322 (2012).
4. Sadler, P. M. & Tai, R. H. The two high-school pillars supporting college science. *Science* **317**, 457–458 (2007).
5. Rose, H. & Betts, J. R. *Math Matters: The Links Between High School Curriculum, College Graduation, And Earnings*. (Public Policy Institute of CA 2001).
6. Goodman, J. The labor of division: Returns to compulsory high school math coursework. *J. Labor Econ.* **37**, 1141–1182 (2019).
7. Joensen, J. S. & Nielsen, H. S. Is there a causal effect of high school math on labor market outcomes? *J. Hum. Resour.* **44**, 171–198 (2009).
8. Crosnoe, R. & Schneider, B. Social capital, information, and socioeconomic disparities in math course work. *Am. J. Educ.* **117**, 79–107 (2010).
9. Bell, A., Chetty, R., Jaravel, X., Petkova, N. & Van Reenen, J. Who becomes an inventor in America? The importance of exposure to innovation. *Q. J. Econ.* **134**, 647–713 (2018).
10. Xie, Y., Fang, M. & Shauman, K. STEM education. *Annu. Rev. Sociol.* **41**, 331–357 (2015).
11. Hamrick, F. A. & Stage, F. K. College predisposition at high-minority enrollment, low-income schools. *Rev. High. Educ.* **27**, 151–168 (2004).
12. Garet, M. S. & DeLany, B. Students, courses, and stratification. *Sociol. Educ.* **61**, 61–77 (1988).
13. Smith-Woolley, E. et al. Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. *npj Sci. Learning* **3**, 3 (2018).
14. Hoyer, E. & Buenrostro, M. Supporting STEM access, equity, and effectiveness: equitable access to rigorous STEAM coursework (2018). https://www.csba.org/-/media/CSBA/Files/GovernanceResources/GovernanceBriefs/2018GovBriefSTEM2.ashx.
15. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
16. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
17. Wray, N. & Visscher, P. Estimating trait heritability. *Nat. Educ.* **1**, 29 (2008).
18. Purcell, S. M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
19. Brown, J., US Department of Education, Dalton, B., Laird, J. & Ifill, N. *Paths through Mathematics and Science: Patterns and Relationships in High School Coursetaking* (Institute of Education Sciences, 2018).
20. Roderick, M. & Camburn, E. Risk and recovery from course failure in the early years of high school. *Am. Educ. Res. J.* **36**, 303–343 (1999).
21. Stevenson, D. L., Schiller, K. S. & Schneider, B. Sequences of opportunities for learning. *Sociol. Educ.* **67**, 184–198 (1994).
22. Boaler, J. & Staples, M. Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record* **110**, 608–645 (2008).
23. Harris, K. M. et al. Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyz115 (2019).
24. Perkins, R. *The High School Transcript Study: A Decade of Change in Curricula and Achievement, 1990–2000* (DIANE Publishing, 2004).
25. Muller, C. et al. *Wave III Education Data: Design and Implementation of the Adolescent Health and Academic Achievement Study* (Carolina Population Center, Chapel Hill, NC UNC-CH, 2007).
26. Trejo, S. et al. Schools as moderators of genetic associations with life course attainments: evidence from the WLS and Add Health. *Sociol. Sci.* **5**, 513–540 (2018).
27. Belsky, D. et al. Genetic analysis of social-class mobility in five longitudinal studies. *Proc. Natl. Acad. Sci.* **115**, E7275–E7284 (2018).
28. Fletcher, J. M. Social interactions and college enrollment: a combined school fixed effects/instrumental variables approach. *Soc. Sci. Res.* **52**, 494–507 (2015).
29. Stevenson, D. L. & Schiller, K. S. State education policies and changing school practices: Evidence from the National Longitudinal Study of Schools, 1980-1993. *Am. J. Educ.* **107**, 261–288 (1999).
30. Weiner, J. Sankey or ribbon plots. R package version 0.6. https://rdrr.io/cran/riverplot/ (2017).
31. de Zeeuw, E. L. et al. The moderating role of SES on genetic differences in educational achievement in the Netherlands. *npj Sci. Learning* **4**, 13 (2019).
32. Hegedus, A. Evaluating the relationships between poverty and school performance. NWEA Research. *NWEA* (2018).
33. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
34. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
35. Tucker-Drob, E. M. Measurement error correction of genome-wide polygenic scores in prediction samples. *bioRxiv* https://doi.org/10.1101/165472 (2017).
36. Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563 (2016).
37. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv* 588020 https://doi.org/10.1101/588020 (2019).
38. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
39. Crosnoe, R. & Muller, C. Family socioeconomic status, peers, and the path to college. *Soc. Probl.* **61**, 602–624 (2014).
40. Wertz, J. et al. Genetics of nurture: A test of the hypothesis that parents' genetics predict their observed caregiving. *Developmental Psychol.* **55**, 1461–1472 (2019).
41. Wertz, J. et al. Using DNA from mothers and children to study parental investment in children's educational attainment. *Child Dev.* https://doi.org/10.1111/cdev.13329 (2019).
42. Armstrong-Carter, E. et al. The earliest origins of genetic nurture: Prenatal environment mediates the association between maternal genetics and child development. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/2usk8 (2019).
43. Domingue, B. W. & Belsky, D. W. The social genome: Current findings and implications for the study of human genetics. *PLoS Genet.* **13**, e1006615 (2017).
44. Koellinger, P. D. & Harden, K. P. Using nature to understand nurture. *Science* **359**, 386–387 (2018).
45. Tucker-Drob, E. M. & Harden, K. P. A behavioral genetic perspective on noncognitive factors and academic achievement. *Genetics, Ethics and Education (Current Perspectives in Social and Behavioral Sciences* 134–158 (Cambridge University Press, 2017).
46. Simpkins, S. D., Davis-Kean, P. E. & Eccles, J. S. Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Dev. Psychol.* **42**, 70 (2006).
47. Tucker-Drob, E. M. & Harden, K. P. Learning motivation mediates gene-by-socioeconomic status interaction on mathematics achievement in early childhood. *Learn. Individ. Differences* **22**, 37–45 (2012).
48. Luo, Y. L., Kovas, Y., Haworth, C. M. & Plomin, R. The etiology of mathematical self-evaluation and mathematics achievement: understanding the relationship using a cross-lagged twin study from ages 9 to 12. *Learn. Individ. Differences* **21**, 710–718 (2011).
49. Branigan, A. R., McCallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. *Soc. Forces* **92**, 109–140 (2013).
50. Cesarini, D. & Visscher, P. M. Genetics and educational attainment. *npj Science of Learning* **2**, 4 (2017).
51. Domingue, B. W., Belsky, D. W., Conley, D., Harris, K. M. & Boardman, J. D. Polygenic Influence on Educational Attainment. *AERA Open* **1**, 1–13 (2015).
52. Rietveld, C. A. et al. Replicability and robustness of genome-wide-association studies for behavioral traits. *Psychological Sci.* **25**, 1975–1986 (2014).
53. Belsky, D. W. & Harden, K. P. Phenotypic annotation: using polygenic scores to translate discoveries from genome-wide association studies from the top down. *Current Directions in Psychological Science*, https://doi.org/10.1177/0963721418807720 (2019).
54. Martschenko, D., Trejo, S. & Domingue, B. W. Genetics and education: recent developments in the context of an ugly history and an uncertain future. *AERA Open* **5**, https://doi.org/10.1177/2332858418810516 (2019).
55. Sokolowski, H. M. & Ansari, D. Understanding the effects of education through the lens of biology. *npj Sci. Learning* **3**, 1–10 (2018).

56. McQueen, M. B. et al. The national longitudinal study of adolescent to adult health (add health) sibling pairs genome-wide data. *Behav. Genet.* **45**, 12–23 (2015).
57. Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
58. Frank, K. A. et al. The social dynamics of mathematics coursetaking in high school. *Am. J. Sociol.* **113**, 1645–1696 (2008).
59. Curtin, T. R., Ingels, S. J., Wu, S. & Heuer, R. *National Education Longitudinal Study of 1988: Base-year to Fourth Follow-up Data File Users Manual*. (US Department of Education, Office of Educational Research and Improvement, 2002).
60. Boardman, J. D., Domingue, B. W. & Fletcher, J. M. How social and genetic factors predict friendship networks. *Proc. Natl Acad. Sci.* **109**, 17377–17381 (2012).
61. Christensen, R. H. B. Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal. R-package version 1–31 (2015).
62. Keller, M. C. Gene × environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol. Psychiatry* **75**, 18–24 (2014).
63. Cameron, A. C. & Miller, D. L. A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* **50**, 317–372 (2015).
64. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc.: Ser. B (Methodol.)* **42**, 109–127 (1980).

ACKNOWLEDGEMENTS

B.W.D is supported by award #96-17-04 from the Russell Sage Foundation and the Ford Foundation. K.P.H., D.W.B., and E.M.T.D. are supported by Jacobs Foundation Research Fellowships. K.P.H, E.M.T.D., and R.C. are Faculty Research Associates of the Population Research Center at the University of Texas at Austin, which is supported by a grant 5-R24-HD042849 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). Research by K.P.H. and E.M.T.D. is further supported by NICHD grant R01-HD083613.

AUTHOR CONTRIBUTIONS

K.P.H. and B.W.D. contributed equally to the project and were jointly responsible for study conception and design. B.W.D. conducted statistical analyses and prepared figures. K.P.H., B.W.D. and D.W.B. drafted the manuscript. J.W.B., R.C., M.M., M.N., E.M.T.D. and K.M.H. provided critical feedback on analyses, results, and drafts of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41539-020-0060-2.

Correspondence and requests for materials should be addressed to K.P.H.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020