Bringing up the bio-datafied child

Bringing up the bio-datafied child: scientific and ethical controversies over computational biology in education

Ben Williamson, Centre for Research in Digital Education, University of Edinburgh, UK, https://orcid.org/0000-0001-9356-3213

Abstract

Scientific advances in genetic analysis have been made possible in recent years by technical developments in computational biology, or bioinformatics. Bioinformatics has opened up the human genome to diverse analyses involving automated laboratory hardware and machine learning algorithms and software. As part of an emerging field of social genomics, recent educational genetics studies using big data have begun to raise challenging findings linking DNA to predicted life outcomes. Bioinformatic technologies and techniques including 'genome-wide association' and 'polygenic scoring' are producing new kinds of genetic biodata and expert knowledge for rethinking the upbringing and education of children. This article takes as an empirical opening a recent epistemic and ethical controversy over the use of biodata to make genetic predictions about educational, socio-economic and life outcomes, detailing the forms of expert knowledge and technologies involved these predictions, and exploring the social and ethical implications of data-intensive bioinformatics.

Keywords: Behavioural genetics, big data, bioinformatics, biology, genoeconomics, sociogenomics

Introduction

The aim of this contribution is to examine a scientific and ethical controversy over the production and use of computer-generated biological data in education-related research. The

controversy has emerged from three converging developments. First, over the past two decades, new 'educational data scientists' have gained disciplinary and professional authority in educational research, and begun undertaking 'big data' studies into the cognitive, neural and embodied substrates of human learning (Williamson 2020). Second, fresh interest in genetic science in education has brought biological accounts of human subjectivity into education research as a route to calling for policy and/or practice innovations (Youdell & Lindley 2018). And third, the biological sciences have become increasingly 'data-intensive' with the development of advanced computational power, data analytics and machine learning in human genomic research (Leonelli 2016). Computational biology, or bioinformatics—the integration of biology and informatics—has become central to data analysis, discovery and knowledge production in the biological sciences, profoundly affecting how human life is examined and understood (Stevens 2013).

At the intersection of educational data science, educational genetics and bioinformatics, researchers have developed methods for creating precise policy-relevant knowledge about the ways that biology impacts on learning and educational outcomes, and begun exploring possibilities for intervention informed by biological knowledge (Gulson & Webb 2018). Recent educational genomics studies have generated new knowledge through bioinformatic analyses of complex associations between DNA and educational outcomes (attainment, test performance, achievement, cognitive and non-cognitive skills), as well as longer-term socio-economic and life outcomes and traits such as intelligence (Martschenko, Trejo & Domingue 2019). The bioinformatics infrastructures, expert practices and methods employed in these tasks, however, remain under-researched from critical social scientific or philosophical perspectives. The empirical focus of this article is ongoing controversy over data-intensive bioinformatic methods in research on education, addressing how bioinformatics experts and technologies are intervening in research practices, what new knowledge is being produced and

how that knowledge is translated into possible (albeit contested) implications for educational practice and policy.

These emerging fields represent a significant shift in the life sciences to incorporate social data in studies of biological processes, described as a novel form of 'social science genetics', 'social genomics' or 'sociogenomics' owing to its use of genetic data to explore social issues normally reserved to social science disciplines (Braudt 2018; Mills & Tropf 2020). Sociogenomics is part of 'an emerging trend: using huge amounts of data and computing power to uncover genetic contributions to complex social traits', which raises important 'ethical and societal risks of acting on such information' (Adam 2019). As sociogenomics studies have proliferated to examine an increasing range of educational factors, these developments stand to reposition education as a datafied 'precision science': a new model of 'precision education' akin to 'precision medicine' in the biomedical field that positions human bodies and life itself as objects of policy and practice interventions (Kuch, Kearnes & Gulson 2020). The contention by certain scientists of sociogenomics that 'precision education' is possible and desirable, using genetic data from bioinformatic analyses to 'personalise' learning processes (Plomin & von Stumm 2018), and the specific scientific and ethical controversies generated by these proposals for predictive DNA-based policy and practice, are the focus of this article. It contributes insights into the digitalization of children and upbringing that is the focus for this special section, illuminating how new bio-datafied configurations of human subjects are being generated by bioinformatics studies related to education and childcare.

Bioinformatics and sociogenomics

Since the sequencing of the human genome—the entire genetic structure of human DNA—was finalised by the Human Genome project in 2003, 'postgenomic' science has advanced considerably, ushering in biotechnological and biomedical innovations while simultaneously

catalysing concern about genetic modification (Reardon 2017). In this postgenomic context, new knowledge about human life is generated from bioinformation and biodata that are analysed in the dry lab of computational hardware and software rather than the wet lab of biological samples and instruments, with profound consequences for how human life is understood, conceptualized and treated (Parry & Greenhough 2018). The field of genetics, born partly out of nineteenth-century innovations in biostatistics, has metamorphosed into bioinformatic forms of analysis, examination, dissection and understanding of life 'with and through computers' (Stevens 2013, 11).

The transformation of biology into the information science of bioinformatics, and the transcoding of organic biological samples into bioinformational data, raises a host of issues regarding the ways biodata are accessed, circulated or capitalised on. The assembling and analysis of big data in computational biology entails radical new divisions of research labour and training for biological scientists in computer programming, data analytics and information infrastructures, as biology has moved from *in vivo* experimentation to *in silico* databases and pattern detection (Leonelli 2018). Bioinformatics does not only signify the physical transformation of human tissue and DNA into machine-readable digital data, but is also implicated in bodily commodification, the emergence of biocapital markets, the expansion of the biotech industry, and questions of ethics, inclusion, exclusion and inequality (Reardon 2017). The act of bringing up digital biodata from human bodies for analysis in the bioinformatics lab is shaped by the bioeconomy of informatic capitalism and the biotech industry, which in turn shapes how human life is conceived and acted upon in fields such as 'precision medicine' and 'personalized healthcare' (Prainsack 2018).

The education and upbringing of children have emerged as key concerns in the context of postgenomic science. Significant debates and controversies have emerged about DNA testing babies to make predictions about later life outcomes, predisposition to disease or other medical

conditions, and the prospects of engineering 'designer children' (Regalado 2019). Postgenomic science has opened up the genetic factors involved in education as a new frontier of knowledge production and potential policy influence too, as 'big biodata' about the human genome have been analysed to uncover genetic associations with learning, intelligence, attainment, and achievement (Gaysina 2016). Education has now become a central concern of the field of sociogenomics, with advocates claiming social scientific analyses must integrate 'data on the genetic material that underlies our human bodies with the study of how those bodies are differentially treated in the social world' (Braudt 2018, n.p.).

Sociogenomics has also become highly politically charged. Conservative commentators have mobilised recent sociogenomic research to dismiss 'left-wing' or 'progressive' explanations for variations and inequalities in educational outcomes, and to support policies that take DNA differences related to intelligence, achievement and attainment into account (Young 2018; Murray 2020). Sociogenomic expertise in education therefore raises ethical concerns about biological determinism and eugenics, reanimating longstanding debates about the genetic inheritance of intelligence and ability (Youdell & Lindley 2018). As such, sociogenomics is a site of controversy in two senses: first, a site of scientific controversy between different perspectives and lab settings, and second, a site of ethical controversy concerned with the scientific measurement, differential treatment, and potential discriminatory practices on children (Comfort 2018).

Bioscience controversies

The analysis in this article builds on science and technology studies (STS), approaching bioinformatics as a combination of social, technical, human, economic and political relations that are structured by certain forms of knowledge, practices, assumptions, and shared goals, and that engender particular ethical ways of thinking about human life and societal order (Rose

2007). Bioinformatics technologies are imprinted with particular ways of understanding human life that emerge both from existing forms of scientific knowledge and from the specific computational and analytical capacities available from their programming (Stevens 2017).

A methodological concern in STS is the study of epistemic and public controversies, or the disagreements, negotiations and compromises that occur in relation to scientific knowledge production, its reception by the public, and its translation into policy (Meloni & Testa 2018). These moments of controversy are important because they demonstrate how scientific knowledge production is thoroughly socially embedded, situated relationally in material and technical infrastructures, and often in tension with ethical judgments, especially where the stakes involve the manipulation of biological life (Prainsack 2017). Controversies also reveal instability in the direction of science, calling analytical attention to the 'social life' of scientific innovation and the contingency of knowledge production.

Methodologically, the article traces a scientific and ethical controversy over sociogenomics through a corpus of texts produced by scientists involved in research about the genetics of educational outcomes. These texts are read interpretively, 'not simply as empirical reports but as evidence of relationships between key sociotechnical elements', and by treating these texts as 'informants' that provide insight into the concepts, practices and debates that characterise the controversy (Perrotta and Selwyn 2019, 4). The sociotechnical elements detailed in the following sections include the scientific organizations and individual experts involved in the controversy over educational sociogenomics; the apparatus of bioinformatics technologies and methods that shape their research findings; and their translation into proposals for policy and/or practice.

Analytically, the key argument is that experts and technologies of sociogenomics are producing novel bioinformatic policy objects—datafied renderings of human populations, lives and

bodies as biodata which are created in such a way to shape implications and proposals for policy and practice interventions. The apparatus of bioinformatics, as in other fields of datafied knowledge production, not only carries authority to 'discover' or 'reveal things' but also to 'bring into being the very objects they are meant to describe and represent' (Ruppert 2018, 19). As biological science has transformed into bioinformatics, 'computers have altered our understanding of "life"', as 'biological objects' have been 'virtualized' as codes, sequences and patterns, and 'databases and algorithms determine what sorts of objects exist' for analysis and knowledge generation (Stevens 2013, 5).

As such, the title 'bringing up the bio-datafied child' refers to (1) how computational biology brings into being new ways of understanding children and predicting their life outcomes based on data analysis of DNA, and (2) the influence of computational biology on the ways parents, educators and policymakers may approach the education and upbringing of children. The claim that individually personalised 'precision education' could be designed based on computational genetic analysis, in particular, carries profound social and ethical consequences for the ways that children may be understood in biological terms (by educators, parents and policymakers), and for how knowledge produced from computational biodata analysis may be mobilised in practice and policy. Though as shown below, the path forward for sociogenomic studies of education faces in two ways as scientists disagree about the appropriate interpretation and use of novel biodata.

Expert lab settings

Two key lab settings have established trajectories for bioinformatics-based educational genomics studies, and are at the centre of the scientific and ethical controversy studied here. The first is a research cluster at the Social, Genetic and Developmental Psychiatry Centre (SGDP) at King's College London, centred on behavioural geneticist Robert Plomin, along

with other collaborators. The other is the Social Science Genetics Association Consortium (SSGAC), an internationally distributed consortium with members in the US, Canada, Australia and Europe, established and led by the genoeconomist Daniel Benjamin of the University of Southern California, Los Angeles. These groups pursue similar research questions and methodologies, and together their studies are part of the emerging subfield of sociogenomics, yet they have begun to diverge sharply in terms of the interpretations and policy and practice implications they derive from the data.

The SGDP group has firmly established itself to produce policy-relevant sociogenomic evidence, based on its research track record in behavioural genetics, molecular genomics, and statistical genetics).[1] In their systematic review of the contribution of behavioural genetics to understanding the associations between genes, cognitive ability and education, Plomin's team argue that: 'Considering both cognitive and noncognitive skills as well as their biological and environmental underpinnings will be fundamental in moving towards a comprehensive, evidence-based model of education' (Malanchini et al. 2020, 229).

A particular understanding of education infuses the SGDP research: (1) drawing on OECD, that 'Educational attainment is a measure of human capital and is indicative of the skills of a population'; (2) 'extant research has identified general cognitive ability as the major source of variation in academic performance, measured as both school achievement and how long people spend in education—i.e. educational attainment'; and (3) 'the observed associations between cognitive ability and education … are rooted in genetic variation' (229). This identifies the importance of human capital development as an international education policy agenda; situates general cognitive ability—intelligence—as a key indicator of human capital, as measured through school achievement and attainment proxies; and claims intelligence has clear genetic underpinnings.

The SSGAC has become an established centre of expertise in polygenic scoring and genoeconomics in education, operating as a 'distributed lab' across many institutions and disciplines in a variety of countries. It was founded in 2011 as a research consortium to test the feasibility of an 'alternative, rigorous, large-data approach to social science genetics' (https://www.thessgac.org/). Its specialism is 'big data' meta-studies involving very large samples of biodata on phenotypes/traits, 'including attitudes, behaviors, economic preferences, and socioeconomic outcomes' (https://www.thessgac.org/phenotypes). It also shares datasets publicly for re-use by other researchers (https://www.thessgac.org/data). Plomin's team used SSGAC data in conjunction with UK exams data to predict school achievement in high-stakes assessments. The SSGAC is not a conventional biological lab that works with 'wet' samples; instead, it compiles digital biodata for analysis seeking to identify social-scientific phenotypes such as behaviours and socioeconomic outcomes.

The SSGAC's most high-profile project is a long-term study of the genetics of educational attainment. The study itself is the accomplishment of a well-funded international team of 80 scientists working in departments of psychology, sociology, behavioural genetics, behavioural science, neurogenomics, economics, biosciences, health sciences, and many others including scientists from the commercial organization 23andMe, the Silicon Valley consumer genetics ancestry company backed by Google (and owner of the world's largest private database of consumer biodata). Its research, then, is distributed across public universities and commercial labs at huge scale and significant cost.

Controversy among the SGDP and SSGAC groups first emerged in 2016 when SSGAC colleagues reported 'divisive' findings in *Nature* (Hayden 2016) showing the very small effect of genes on educational attainment (years spent in education). The key finding was that DNA analysis could be used to predict a small percentage of the variation in educational attainment, on average rather than at the individual level, with the secondary finding that attainment

worked as an accurate proxy of intelligence because of its associations with cognitive performance test scores. As reported in the media, however, Plomin interpreted the findings as a 'tipping point' for future studies 'to predict genetic strengths and weaknesses for individuals', which he argued could be used to 'move education closer to "personalised learning" rather than continuing to assume that a one-size-fits-all national curriculum works equally well for everyone' (Sample 2016). He also argued such studies could pave the way to 'predictive genetics for traits such as how well children perform on standardized tests', although the SSGAC termed such proposals 'irresponsible' and scientifically implausible (Hayden 2016).

These lab settings—and their disagreement—illuminate how the labour of producing new knowledge related to educational outcomes has moved to new and competing sites, where the disciplinary expertise and assumptions of fields of behavioural genetics, molecular genomics and genoeconomics differentially structure the kinds of scientific endeavours that take place and shape what interpretations and knowledge are produced. This process of knowledge production, however, also relies on complex sociotechnical bioinformatics infrastructure for its realization.

Biodatafied knowledge production

A number of key bioinformatics innovations underpin the capacity of sociogenomics labs to produce knowledge. For the SGDP, 'molecular genetic research, particularly recent cutting-edge advances in DNA-based methods, has furthered our knowledge and understanding of cognitive ability, academic performance and their association' (Malanchini et al. 2020, 229-230). They describe these cutting edge methods as part of a 'technological advance' that 'enabled an atheoretical approach to identify associations across the genome', and which, through 'statistical power', have led to 'increasingly more insight into the molecular genetic architecture of cognitive ability and academic performance' (Malanchini et al. 2020, 235). This

emphasis on technologically-advanced and atheoretical analysis of molecular genetic architecture is typical of the shift to 'data-centric' rather than theory-driven biology that utilises bioinformatics for knowledge discovery from vast datasets (Leonelli 2016). It also exemplifies a 'molecular style of thinking' in contemporary biology that conceives of human life at the scale of molecular entities and their interactions, which 'is fabricating a new way of understanding life itself':

> Molecular genomics has depended on the invention of a whole range of technologies for decomposing, anatomizing, manipulating, amplifying and reproducing vitality at this molecular level. … [T]hese techniques opened 'the gene' to knowledge and technique at the molecular level. (Rose 2007, 13-14)

Bioinformatics has been central to the transformation of human biology into molecular genomics. The field of sociogenomics has now brought these techniques and a molecular gaze to bear on educational questions and the anatomization of the molecular genetic architectures underpinning student outcomes. Two technologies in particular are integral to the knowledge production techniques of the SGDP and SSGAC: microarrays and biobanks.

Microarrays

Molecular sociogenomic studies require hardware and software capable of capturing and identifying minute DNA variations scattered across different regions of the genome. Microarrays are matchbox-sized glass slides, termed 'labs-on-a-chip', imprinted with DNA fragments. When a dissolved sample of a patient's or client's DNA flows across the slide, it binds with any complementary fragments. It can then be analysed in a microarray scanner or an automated 'laboratory robot' capable of analysing thousands of arrays, which is necessary for genotyping an individual in terms of how they differ from a group or population (Kragh-Furbo et al. 2016). A huge industry of bioinformatics infrastructure suppliers has developed to

support microarray-based genomics, as science has hybridised with business and technology in contemporary data-centred biology (Reardon 2017).

Genomic science has labelled the variations identified by microarrays as single nucleotide polymorphisms (SNPs), tiny building blocks in human DNA. Each individual person's genome has 4-5 million SNPs, and across large populations hundreds of millions of SNPs have been identified. Mapping patterns among these variations, especially across large populations, has become the basis for Genome-Wide Association Studies (GWAS). These molecular genomics studies resist simplistic explanations that any single gene or cluster of genes is responsible for disease risk, susceptibility to inherited health problems, or behavioural trait, instead seeking out complex architectural associations across millions of SNPs (Plomin 2018).

By aggregating and weighting identified SNPs, scientists produce genome-wide polygenic scores, or predictors of a particular phenotype, trait or outcome based on patterns found in the analysed SNP sample. This is a complex statistical and computational procedure, involving bioinformatics hardware, software packages, algorithms, and machine learning methods. Associations between SNPs and traits or diseases 'are highly numerically leveraged', relying on 'many crosshatched layers of data, inference, modelling, and testing', with the DNA microarray 'effectively running thousands or millions of experiments in parallel':

> The toolboxes of computer science and statistics have been effectively ransacked for pattern recognition, data mining, and machine learning techniques that can detect the associations between hundred and thousands of SNPs. (Kragh-Furbo et al. 2016, 11)

Microarray-based studies of gene interactions such as genome-wide association studies of SNPs and polygenic scoring need to be understood as a complex accumulation of scientific knowledges, lab settings and experimental practices, computer science and statistical

techniques, business logics and marketisation, biotechnology devices and automated infrastructures, and models and inferences.

Sociogenomics has embraced microarrays and GWAS methods, especially the mapping of SNP patterns that are associated with educational outcomes. Educational attainment was among the first human cognitive phenotypes to be studied using genome-wide association, SNP and polygenic scoring methods. Robert Plomin's lab in particular has mobilised microarrays, GWAS, SNP detection and polygenic scoring to support its vision of an 'evidence-based model of learning' that integrates 'genetic prediction into personalized approaches to learning and interventions' (Malanchini et al. 2020, 241). As the SGDP lab acknowledge, these new insights into genotype-phenotype associations and prediction in education have only become possible through cutting edge techniques such as SNP microarrays and 'laboratory robots' armed with lasers and digital optics for scanning the chips. Microarray technologies are not just passive tools of biological discovery but 'characterized by the leveraging and assembling of a plurality of forms, standards, knowledge, technologies, and practice', and they depend on 'computer and statistical sciences to recognize patterns of variations and to highlight possible associations. As commodities, microarrays build on molecular biology and genomics, robotics and automation' (Kragh-Furbo et al. 2016, 21). In the polygenic molecular genetics practised at labs such as the SGDP, human biology is reconceived through the analytical gaze of laboratory robots and the toolboxes of the computer, statistical and data sciences.

Biobanking

A second key bioinformatics technology for sociogenomic studies is biobanks, which have been integral particularly to the GWAS research, polygenic scoring and knowledge production of the SSGAC. SSGAC's first GWAS of educational attainment, reported in 2013 and based on a sample of 126,500 individuals, identified three SNPs associated with just 2% of the variation in educational attainment, with a follow-up 2016 of a sample of 300,000 identifying

74 SNPs that could explain 3.2% of the variation. Then in 2018, with an updated sample of 1.1 million, the SSGAC team reported more than 1,200 SNPs that together accounted for 11–13% of the variation in years spent in school—which it claimed explains approximately as much variation in time in education as family socio-economic status.

The paper, 'Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment' (Lee et al. 2018), reported that genetic variants involved in educational attainment particularly included those involved in brain-development processes and the formation of neuronal connections in foetuses and newborns. These early developmental neuro-biological factors, the SSGAC claimed, influence psychological development, which in turn affects how long people continue at school, and continues to exert lifelong influences on socio-economic life outcomes. As such, through polygenic scoring, the SSGAC claims to have established the influence of biology on factors normally measured in social or economic terms, such as educational attainment and occupational status.

Its biodata sample of over a million people was from two 'biobank' sources: the UK Biobank, a huge open access health resource based on a living population of over 500,000 volunteer participants; and the private biobank of 23andMe, the consumer genetics company offering health and ancestry services for profit. By employing the services of 23andMe both as a biobank supplier and as a research partner, the SSGAC has embedded itself as an organization in the infrastructure of 'bioinformatic capitalism' (Reardon 2017). As Reardon (2017, 124-125) notes, 23andMe is 'powered by Google's money and algorithmic prowess', and has helped transform human genetics into 'genomics 2.0' by bringing social media strategies to the field of human genomics, and by vastly increasing both the number of subjects and the amount of data collected about them. The SSGAC has harnessed the power of 23andMe's private biobank as a source of new knowledge production, in the process making this new form of sociogenomic educational research into a new front of the bio-economy. Although the SSGAC

has not conducted polygenic scoring on any samples of children, Plomin (2018) foresees the potential for 23andMe to do so.

Crucially, these exercises in data-mining biobanks for predictive value assume that biobanks provide unmediated insight into genotype-phenotype dynamics. Critical science studies of biobanks, on the contrary, contend that the digitalised samples of biobanks 'are not stable realities, defined once and for all according to fixed and essential characteristics: they are always evolving, formatted by sociotechnical processes that make them shift from the status of living entities—linked to the bodies of the people from which they come from—to "things" or "bio-objects"' (Milanovic, Merleau-Ponty & Pitrou 2018, 285). In biobanks, 'human biological resources are fabricated through the interweaving of a technical processes and not "natural" at all' [sic] (287), but the result of scientific and technical practices of extraction, standardisation, classification, conservation, circulation, commercialisation and commodification.

Microarrays and biobanks exemplify the ways that bioinformatics perform particular operations to define and classify living bodies. They are not passive intermediaries through which scientific expertise can penetrate embodied molecular genetic architectures, but actively mediate the possible meanings available through the intertwined sociotechnical apparatus of computer and data sciences, bioinformatics businesses and biological entrepreneurship. Just as 'precision medicine' and 'postgenomic' sciences are understood as a combination of business, biology and technology, the introduction of sociogenomics into educational knowledge production brings the logics of the commercial bioeconomy and the practices of Silicon Valley data mining to bear on analysing and predicting phenotypes such as educational outcomes and their molecular genotypical associations.

Biodata translations

The third sociotechnical element of the new sociogenomics is the diverging ways in which they have begun translating new research findings into proposals for educational policy and practice. Two particular knowledge claims are detectable: for the SSGAC, that polygenic scores can be used as indicators of socio-economic outcomes, and for the SGDP, that polygenic scores may be used for personalised precision education.

Genoeconomics

The SSGAC, notably, is guided by 'genoeconomics'. Its 'research philosophy' is set out in two papers: 'Molecular genetics and economics' (Beauchamp et al. 2011) and 'The promises and pitfalls of genoeconomics' (Benjamin et al. 2012). The use of molecular genetic information to study economic behaviours and outcomes emerged originally from behavioural genetics, where Plomin is a leading figure, and moved towards the combination of genotype data with socioeconomic datasets, enabling genoeconomists to 'focus on the collection of additional, standardized socioeconomic variables in samples that already contain genotyped data. Over time, adding genotype information to existing socioeconomic datasets should provide an important new source of data to economists interested in studying the molecular genetic associates of economic behaviors and outcomes' (Beauchamp et al. 2011, 24). The molecular genetics approach to economics, or genoeconomics, sees 'economic outcomes and preferences … as heritable as many medical conditions and personality traits' and seeks 'to find associations between genetic variation and economic behavior' through 'pooling datasets', using 'statistical techniques that exploit the greater information content of many genes jointly', and 'focusing on economically relevant traits that are most proximate to known biological mechanisms' (Benjamin et al. 2012, 627). As such, its GWAS consist of data-intensive bioinformatics infrastructure and methods for identifying associations between biological genotypes and behavioural or physical phenotypes, with polygenic scoring increasingly used to predict socio-economic outcomes.

What translations into policy-relevant knowledge follow from this massive sociogenomics study of educational attainment and socio-economic life outcomes? Although the SSGAC is careful to report no immediate policy implications of its findings, one of its founding papers speculates that 'knowledge of the biological mechanisms' underpinning socioeconomic outcomes 'might suggest additional policies or interventions that had previously not been anticipated', such as the use of molecular genetic data to 'predict which individuals are at high "genetic risk" for adverse outcomes' (Beauchamp et al. 2011, 25). It also suggests 'molecular genetic data may prove helpful in understanding variation in policy response across individuals' (27). As such, despite claims to policy agnosticism, the SSGAC was founded upon aspirations to use molecular genetic data, in the shape of polygenic scores, to make individual-level analyses and produce policy-relevant findings.

Their educational attainment study anticipates the further use of biobank data as sources for socio-economic analysis and potential policy intervention, particularly on those individuals who could be categorised as at 'genetic risk'. Taking educational attainment as a socio-economic correlate might produce policy-relevant insights into how such genetic risks based on polygenic scores—or the opposite, genetic potential—can that can be either minimised or maximised. The study also highlights the value of genoeconomics to labour market forecasting, where molecular genomics analysis and polygenic scoring are taken as proxies for future distributions of occupational status and income. It anticipates the use of predictive polygenic scoring for human capital calculation, and using precise molecular genetic information about individuals and collectives for making and testing policy interventions designed to optimise economic outcomes.

Precision education

Plomin's book *Blueprint* provides a clear example of how new molecular and computational understanding of human genomic architectures can translate into policy-relevant knowledge.

Favouring a model of DNA as a biological 'blueprint' that can be detected through polygenic scoring, Plomin (2018) specifically advocates a model of 'precision education' based on tailoring education to students' DNA. Plomin's arguments for personalised, precision education are based on GWAS aggregating the SNPs associated with educational attainment, achievement and cognitive ability. He argues that the results of such studies demonstrate that DNA is 'the best predictor we have of a child's years of education, even better than the environmental effect of family socioeconomic status' (129).

In the version of 'personal genomics' that Plomin promotes, polygenic scores are 'fortune tellers' used to predict a psychological trait such as intelligence or school achievement, and to 'promote promise' (135). Utilising data from the SSGAC, the results, he argues, show that polygenic scores are powerfully predictive of not just years in school, but also of test achievement and intelligence too. 'So if all you know about people is their DNA, you can indeed predict their school achievement', Plomin claims, although he also admits that 'polygenic scores are only probabilistic predictors' which do not accurately capture 'individual differences' (158). As another sociogenomics researcher—and SSGAC advisory board member—puts it: 'Today these polygenic scores—as they are called—are noisy predictors. But rapid progress is being made such that soon a bit of saliva or blood from a newborn will be able to capture her full genetic potential for educational attainment' (Conley 2017).

Building on this promissory quality, Plomin claims polygenic scores are unique in identifying causal links between DNA and psychological traits and effects such as school attainment, achievement and intelligence. 'DNA differences captured by the polygenic score cause differences between children in their school achievement,' though it 'does not tell us about the brain, behavioural or environmental pathways by which the polygenic score affects the trait' (Plomin 2018, 162). He even goes on to argue that since 'polygenic score predictions are specific to an individual' (163), then it follows that 'polygenic scores are key for personalized

learning, as they predict pupils' profiles of strengths and weaknesses, which offers the possibility to intervene early to prevent problems and promote promise' (181).

Plomin suggests parents may in the first instance take up the opportunities of purchasing polygenic scores for their children, as a way of informing their choices about education and schooling based on 'getting a glimpse of their children's individuality—their strengths and weaknesses, their personalities and their interests. This information might help parents to try to maximize their children's strengths and minimize their weaknesses' (178). Consumer ancestry companies such as 23andMe could begin offering polygenic scores to parents in order to 'arm themselves with their child's genetic blueprint' (178). This is already happening, as an SSGAC-affiliated report states, since '23andMe customers have the possibility to download their own genetic data to inspect it themselves or to share it further, for example with websites such as dna.land, which calculates a polygenic score value for educational attainment for its customers' (Angers et al. 2019, 28).

The promissory quality of precision sciences, which have expanded beyond the biomedical domain, is based on the anticipation that new forms of knowledge—and commercial opportunity—will emerge from the compilation of diverse and ever-larger forms of data (Kuch et al. 2020, 3). Indeed, Plomin's lab has set out a research agenda to increase the scale and longitudinal reach of sociogenomic studies of education and to embed genotype analysis into predictive precision education:

> identifying the specific environmental and biological processes that lead from genetic predisposition to observed variation in cognition and education remains one of the major challenges for future research. Understanding the mechanisms by which genotypes become phenotypes is likely to lead to advances in more effective personalised approaches to learning. (Malanchini et al. 2020, 242).

In sum, then, the Social, Genetic and Developmental Psychiatry Centre has mobilised bioinformatics knowledge, based on big data about education and methods of polygenic scoring, to promote a model of evidence-based personalised learning modelled on precision medicine in the biomedical field.

The SSGAC fundamentally rejects Plomin's precision education vision. While its results show on average that higher polygenic scores are associated with longer educational attainment, huge variation exists between individuals, making individual-level polygenic prediction unfeasible. Key European members of the SSGAC also wrote a report for the European Union on the policy implications of genome-wide association, polygenic scores and social science genetics, specifically arguing that 'it is probably not a good idea to base educational policies, recommendations, tracks, and experiences on a biomarker that carries very little information for any specific individual' (Angers et al. 2019, 25).

This ongoing disagreement, first emerging in 2016 when Plomin mobilised SSGAC findings to support his argument for precision education (Hayden 2016), illustrates the bidirectionality of sociogenomic research in related to education. Viewed as a scientific controversy, it highlights how sociogenomics remains characterised by tensions and frictions—even regarding the same datasets—and the ways that 'competing epistemologies' are informed by 'imaginaries and visions' as much as by the available data and evidence (Meloni & Testa 2018, 198). Far from consensus about sociogenomics, it remains uncertain, speculative and the subject of dispute. As with the field of postgenomic science more broadly, 'it is in this mismatch between what is established and what is at present a source of heated scientific dispute that speculative assumptions, inflated discourses and enthusiastic media promotion … are likely to find fertile ground' (201).

Social and ethical implications

The possibilities of social genomics are exemplified by the development of behavioural genetics and genoeconomics in educational research and knowledge production. The new knowledge of education produced by these forms of sociogenomics depend on the apparatus of bioinformatics—hardware, software, experts, knowledge, and laboratory settings—for the generation of novel biodata, with significant potential consequences for how human subjects are known, understood and acted upon through various practices or policy interventions.

As yet the translation of biodata into knowledge remains a contingent, contested and controversial enterprise. The controversy mapped in this article demonstrates that sociogenomics remains a science-in-the-making rather than a settled field. Computerised knowledge and epistemic expertise are opening up new sites of conflict over epistemic significance, exacerbated by speculation and hype that sociogenomics may revolutionise the study of socioeconomic outcomes such as educational attainment and achievement, or even contribute to DNA-based precision education. It appears to face two ways at once: towards a future of precision-targeted services modelled on precision medicine; or towards a socioeconomic forecasting based on large population samples. Either way, the uptake of sociogenomics methods, knowledge and bioinformatics techniques raises a series of key social implications and ethical controversies.

The first is that sociogenomics operates at a distance from the bodies it analyses. Most obviously, it involves analysis of biodata samples collected from adult volunteers or paying customers of biobanks. These data, despite the veneer of objectivity imputed to ‘big data’ and bioinformatic styles of analysis, are highly partial. Data collected from the UK Biobank, for example, consist of a sample of volunteer adults aged over 50 of white European descent, while the 23andMe data used to construct the SSGAC dataset reflects the company’s customer base rather than a fully representative population sample. The polygenic scores so produced can

only then be said to represent very specific socio-demographic and ethnic selections which, as with precision medicine, raises clear tensions regarding social inclusion and exclusion (Prainsack 2017).

Moreover, knowledge produced through bioinformatics is mediated through many sociotechnical layers of biobanks, microarrays, robot scanners, polygenic scoring and GWAS techniques and technologies, further distantiating the living body from its analysis. Stevens (2017, 153) notes, 'data-inside-computers' is created 'so that it can take on distinct forms and be manipulated in ways that are not possible without a computer'. Bioinformaticians are involved in different knowledge-making practices derived substantially from the development of algorithms and analytics in the technology industry rather than biology itself, and 'the quantities of data processed by computers, and the algorithms needed to deal with them, make a qualitatively different kind of knowledge' (172). In the vision of precision education based on 'personal genomics', children themselves would become biodata transmitters of their own DNA 'blueprint', perhaps from before birth, with their 'genetic futures' foretold by predictive polygenic scores (Plomin 2018, 187). These genetic futures, however, are fabrications of the bioinformatics infrastructure of microarrays, biobanks, polygenic scoring algorithms and the epistemic practices of behavioural geneticists and genoeconomists, which act as powerful mediators between embodied subjects and their representation as calculated biodata objects.

These knowledge-making practices of bioinformatics produce novel conceptualizations of human vitality. In bioinformatics human bodies are understood as 'networked and calculable' bio-objects, taking on 'the characteristics of the cyberinfrastructure that they are connected to, in the sense that the analytical logic of informatics pervades the view on biological and more specifically genomic processes and thus also on human bodies' (van Baren Nawrocka, Consioli & Zwart 2019, 3). Sociogenomics analyses concentrate on bio-datafied subjects that exist in a different epistemic and ontological domain—the database—than the embodied subjects from

which the original DNA samples were taken. They characterise bodies in informational terms, as biodigital bodies constituted of computational codes, sequences, networks and calculations rather than as embodied subjectivities (Parry & Greenhough 2018). Bioinformatics 'brings up' an informatised body for analysis, and its diagnoses and predictions might (ideally for precision education promoters) inform how parents and teachers approach the education and upbringing of children too.

The informatisation of the body in sociogenomics also contributes to the creation of new 'biosocialities'. Biosociality refers to how people relate to themselves as 'molecular-genetic identities', to form collective forms of identification around genetic categories, and to negotiate and make ethical calculations in the light of knowledge about their genetic risks or susceptibilities (Rose 2007, 126). With sociogenomics, a new kind of digitally-defined biosociality is emerging, whereby polygenic scores might become a source of both individual and collective identification, based on one's identification as a subject that has been rendered discernible in patterns and polygenic associations. New digital biosocialities are constructed as genetic collectives categorised by their polygenic scores and 'treated' on that basis. As genoeconomics shows, genome-wide polygenic scoring can even be used as a way of predicting human capital, while precision education advocates argue polygenic scores are useful for parents and teachers to make calculations about the life outcomes of the children in their care, and to act on that knowledge accordingly.

Finally, sociogenomics potentially represents an emerging hereditarian social science with historical links to the biostatistics of the nineteenth century (Porter 2018). Although field experts dispute criticisms that their studies support genetic determinism and caution against simplified policy uptake (Mills & Tropf 2020), sociogenomics findings are circulated in the press and social media by influential high-profile conservative commentators, as well as by extreme alt-right racist groups, reanimating longstanding debates about eugenics and

discrimination stemming from the political uses of intelligence data (Comfort, 2018). Translated into a political science of population prediction, polygenic predictions risks reducing social-structural or environmental causes to deterministic genetic explanations. As polygenic scores for education enter mainstream policy debates, these 'slippery genetic predictions could turn people's attention away from other things that influence how children do in school and beyond—things like their family's wealth, the stress in their neighborhoods, the quality of the schools themselves' (Zimmer 2018). A further risk is polygenic scores being used for discriminatory interventions based on one's calculated biological fortunes—a new 'Eugenics 2.0' based on predictions of a child's genetic educational potential (Regalado 2017).

Conclusion

This initial analysis shows how bioinformatics are implicated in significant changes to how knowledge about educational outcomes are produced, and how these bioinformatics apparatuses are imprinted with particular disciplinary assumptions, speculations and disagreements from sociogenomic fields of behavioural genetics and genoeconomics. This combination of expert knowledge and bioinformatics infrastructure 'brings up' a particular configuration of human life as a code, network or architecture of complex associations only visible through computation, that is, a new form of bioinformational knowledge that stands to affect the ways educators and parents might approach the upbringing of children.

The utilisation of bioinformatics in sociogenomic analyses of education ushers genetic explanations back into debates about bringing up children. New kinds of educational data scientists who move across biology and informatics have begun to conduct novel social genomics analyses and produce new understandings of 'educated' human bodies. The emerging authority of educational data science does not merely 'reveal' objective insights into the biological substrates of learning and educational outcomes, but generates new

understandings shaped by the bioinformatics apparatus required to turn bodies into biodata objects and then translate the data into policy-relevant knowledge.

As disagreement between the 'labs' identified in this article demonstrates, the use of molecular genomic data in knowledge production is not a settled matter. While some propose a 'precision education' model based on individual-level polygenic scoring of children, most extant sociogenomic research does not deliver the individualised findings to support it. Others are building new knowledge about the associations between DNA, education and socio-economic outcomes based on extensive datasets gathered from biobanks. In both cases, the objective is to produce 'precision' in the understanding and prediction of the biological substrates of educational, socio-economic and life outcomes. Mapping the scientific controversy has illuminated some urgent emerging ethical controversies, including the potential for sociogenomic data and predictions to be used for discriminatory practices and policies.

The article has added an important dimension to recent discussions of the impacts of digitization and datafication on children, schools and families. Sociogenomics and precision education proposals appear to open up the possibility for increasing bio-datafication of human subjects, and for forms of digitally-enabled biological prediction that are aimed at securing particular social, economic and personal futures. Opening up sociogenomics in terms of its internal controversies in this article enables us to begin to grasp what kinds of technical, industrial, epistemological and ethical challenges remain to be confronted as educational issues are increasingly treated as complex biological problems that only computers and algorithms can diagnose and, perhaps, solve.

References

Adam, D. 2019. “The promise and peril of the new science of social genomics.” *Nature* 21 October: https://www.nature.com/articles/d41586-019-03171-6.

Angers, A., Kagkli, D., Koellinger, P., Petrillo, M., Querci, M., Raffael, B. and Van Den Eede, G. 2019. “Genome-wide association studies, polygenic scores and social science genetics: overview and policy implications.” Luxembourg: Publications Office of the European Union.

Ball, P. 2018. “The IQ trap: how the study of genetics could transform education.” *New Statesman*, 16 April: https://www.newstatesman.com/2018/04/iq-trap-how-study-genetics-could-transform-education.

Beauchamp, J. et al. 2011. “Molecular genetics and economics.” *Journal of Economic Perspectives* 25 (4): 57-82.

Benjamin, D.J., Cesarini, D., Chabris, C.F. et al. 2012. “The Promises and Pitfalls of Genoeconomics.” *Annual Review of Economics* 4: 627–662.

Braudt, D.B. 2018. “Sociogenomics in the 21st century: an introduction to the history and potential of genetically-informed social science.” *Sociology Compass* 12(10): https://doi.org/10.1111/soc4.12626.

Comfort, N. 2018. “Sociogenomics is opening a new door to eugenics.” *MIT Technology Review*, 23 October: https://www.technologyreview.com/2018/10/23/139420/sociogenomics-is-opening-a-new-door-to-eugenics/.

Conley. D. 2017. “What's Your Polygenic Score?” *Scientific American*, 13 March: https://blogs.scientificamerican.com/guest-blog/whats-your-polygenic-score/.

Gaysina, D. 2016. “How genetics could help future learners unlock hidden potential.” *The Conversation*, 15 November: https://theconversation.com/how-genetics-could-help-future-learners-unlock-hidden-potential-68254.

Gulson, K. & Webb, P.T. 2018. "'Life' and education policy: intervention, augmentation and computation." *Discourse: Studies in the cultural politics of education* 39(2): 276–291.

Hayden, E.C. 2016. "Gene variants linked to success at school prove divisive." *Nature*, 11 May: https://www.nature.com/news/gene-variants-linked-to-success-at-school-prove-divisive-1.19882.

Kragh-Furbo, M., Mackenzie, A., Mort, M. and Roberts, C. 2016. "Do biosensors biomedicalize? Sites of negotiation in DNA-based biosensing data practices." In Nafus, D. (ed.) *Quantified: Biosensing technologies in everyday life*, 5-26. Cambridge, MA: MIT Press.

Kuch, D., Kearnes, M. & Gulson, K. 2020. "The promise of precision: datafication in medicine, agriculture and education." *Policy Studies*: https://doi.org/10.1080/01442872.2020.1724384.

Lee, J.J., Wedow, R., Okbay, A., et al. 2018. "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals." *Nature Genetics 50*, 1112–1121.

Leonelli, S. 2016. *Data-Centric Biology: A philosophical study*. London: University of Chicago Press.

Leonelli, S. 2018. "Assembling biomedical big data." In Meloni, M., Cromby, J., Fitzgerald, D. & Lloyd, S. (eds). *The Palgrave Handbook of Biology and Society*, 317-337. London: Palgrave Macmillan.

Malanchini, M., Rimfield, K., Allegrini, A.G., Ritchie, S.J. & Plomin, R. 2020. "Cognitive ability and education: How behavioural genetic research has advanced our knowledge and understanding of their association." *Neuroscience and Biobehavioral Reviews* 111: 229–245.

Martschenko, D., Trejo, S. & Domingue, B.W. 2019. “Genetics and education: Recent developments in the context of an ugly history and an uncertain future.” *AERA Open* 5(1): 1-15.

Meloni, M. & Testa, G. 2018. “Scrutinizing the epigenetics revolution.” In Meloni, M., Cromby, J., Fitzgerald, D. & Lloyd, S. (eds). *The Palgrave Handbook of Biology and* Society (191-226). London: Palgrave Macmillan.

Milanovic, F., Merleau-Ponty, N. & Pitrou, P. 2018. “Biobanks and the reconfiguration of the living.” *New Genetics and Society* 37(4): 285-295.

Mills, M.C. & Tropf, F.C. 2020. “Sociology, Genetics, and the Coming of Age of Sociogenomics.” *Annual Review of Sociology* 46: 553-581

Murray, C. 2020. “Genetics Will Revolutionize Social Science.” *Wall Street Journal*, 27 January: https://www.wsj.com/articles/genetics-will-revolutionize-social-science-11580169106

Parry, B. & Greenhough, B. 2018. *Bioinformation*. Cambridge: Polity.

Perrotta, C. and Selwyn, N. 2019. “Deep learning goes to school: Toward a relational understanding of AI in education.” *Learning, Media and Technology*: https://doi.org/10.1080/17439884.2020.1686017.

Plomin, R. & von Stumm, S. 2018. “The new genetics of intelligence.” *Nature Reviews Genetics*. doi:10.1038/nrg.2017.104.

Plomin, R. 2018. *Blueprint: How DNA makes us who we are*. London: Allan Lane.

Porter, T. 2018. *Genetics in the Madhouse: The unknown history of human heredity*. Oxford: Princeton University Press.

Prainsack, B. 2017. *Personalized Medicine: Empowered patients in the 21st century?* New York: New York University Press.

Reardon, J. 2017. *The postgenomic condition: Ethics, justice, and knowledge after the genome.* Chicago: University of Chicago Press.

Regalado, A. 2017. "Eugenics 2.0: We're at the dawn of choosing embryos by health, height, and more." *MIT Technology Review*, 1 November: https://www.technologyreview.com/2017/11/01/105176/eugenics-20-were-at-the-dawn-of-choosing-embryos-by-health-height-and-more/.

Regalado, A. 2019. "The world's first Gattaca baby tests are finally here." *MIT Technology Review*, 8 November: https://www.technologyreview.com/s/614690/polygenic-score-ivf-embryo-dna-tests-genomic-prediction-gattaca/.

Rose, N. 2007. *The Politics of Life Itself: Biomedicine, power and subjectivity in the twenty-first century*. Princeton, NJ: Princeton University Press.

Ruppert, E. 2018. *Sociotechnical Imaginaries of Different Data Futures: An experiment in citizen data*. Rotterdam: Erasmus University Rotterdam.

Sample, I. 2016. "Genes that influence how long you stay in education uncovered by study." *The Guardian*, 11 May: https://www.theguardian.com/science/2016/may/11/genes-that-influence-how-long-you-stay-in-education-uncovered-by-study.

Stevens, H. 2013. *Life Out of Sequence: A data-driven history of bioinformatics*. London: University of Chicago Press.

Stevens, H. 2017. "A Feeling for the algorithm: Working knowledge and big data in biology." *Osiris* 32, 151-174.

van Baren-Nawrocka, J., Consoli, L. & Zwart, H. 2019. “Calculable bodies: Analysing the enactment of bodies in bioinformatics.” *BioSocieties*: https://doi.org/10.1057/s41292-019-00143-x.

Williamson, B. 2020. “New digital laboratories of experimental knowledge production: Artificial intelligence and educational research.” *London Review of Education* 18(2): 209-220.

Youdell, D. & Lindley, M. 2018. *Biosocial Education: The social and biological entanglements of learning.* London: Routledge.

Young, T. 2018. “The left is heading for a reckoning with the new genetics.” *The Spectator*, 23 April: https://www.spectator.co.uk/article/the-left-is-heading-for-a-reckoning-with-the-new-genetics.

Zimmer, C. 2018. “Genetic Intelligence Tests Are Next to Worthless.” *The Atlantic*, 29 May: https://www.theatlantic.com/science/archive/2018/05/genetic-intelligence-tests-are-next-to-worthless/561392/.

[1] https://www.kcl.ac.uk/psychology-systems-sciences/about/departments/social-genetic-and-developmental-psychiatry