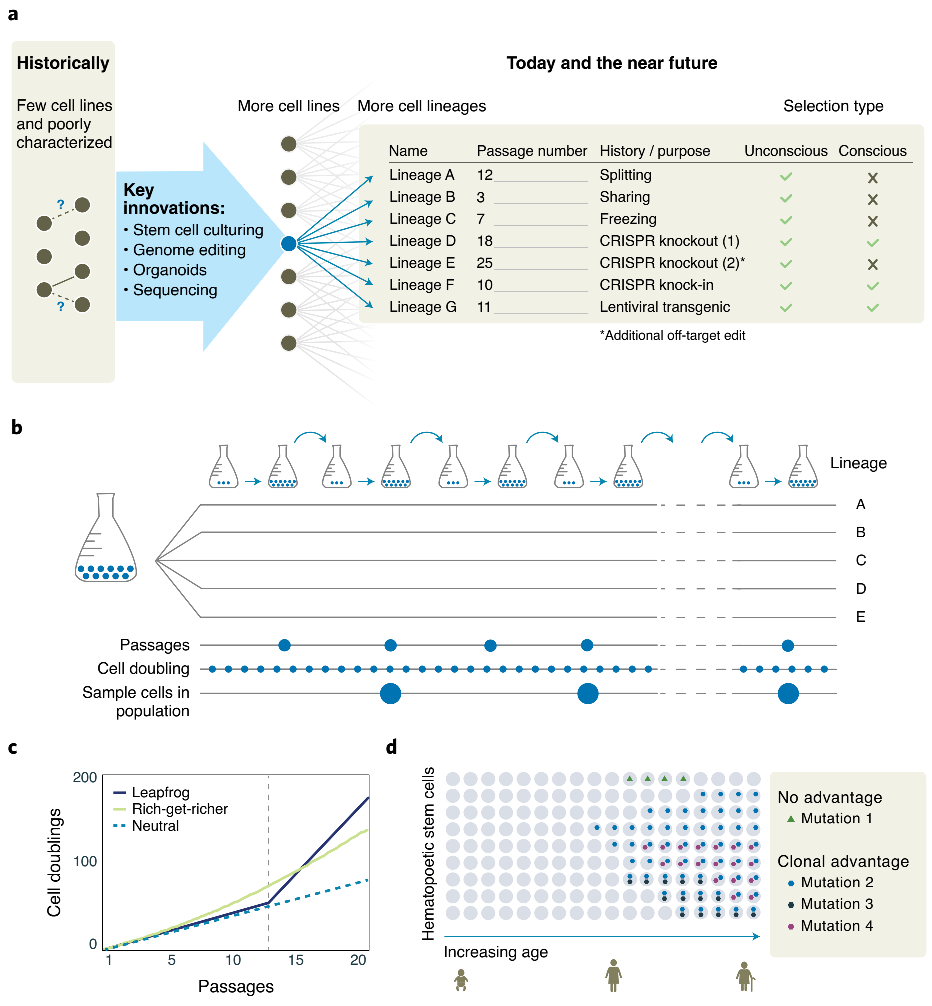
Tracking cell lineages to improve research reproducibility



Name	Passage number	History / purpose	Unconscious	Conscious
Lineage A	12	Splitting	✓	×
Lineage B	3	Sharing	✓	×
Lineage C	7	Freezing	✓	×
Lineage D	18	CRISPR knockout (1)	✓	✓
Lineage E	25	CRISPR knockout (2)*	✓	×
Lineage F	10	CRISPR knock-in	✓	✓
Lineage G	11	Lentiviral transgenic	✓	✓

Fig. 1 | Genome reading and writing technologies have contributed to an expansion in the number of cell lineages and advanced methods to characterize and track these lineages. a, Historically, a limited number of unique cell lines were used. Lineage tracking was minimal, resulting in a knowledge vacuum on cell line provenance (left). Biomedicine requires ever-larger cell line panels representing more patients (gray circles), tissues and diseases. For each cell line, lineages diverge during in vitro maintenance and manipulation (blue arrows). Genetic differences may be consciously engineered, but also occur stochastically (right). The latter are unconsciously influenced by laboratory practices including passaging. **b**, Evolution experiments with clonal microorganisms involve parallel propagation of initially identical cell populations. Each is passaged and sequenced regularly to evaluate stochastic genetic and phenotypic divergence (lineages A–E). **c**, Evolution experiments may show 'rich-get-richer' effects, where fitness benefits compound over time, and 'leapfrog' events, in which lineages with beneficial genetic changes outgrow ones without[7]. Rich-get-richer dynamics were modeled using 10% fitness gains per five doublings and a passage-13 leapfrog event with a fourfold doubling-speed increase. **d**, Somatic cells commonly accumulate stochastic genetic changes over human lifetimes, forming distinct cell lineages by middle age. Although most changes are deleterious or nearly neutral for individual cells, some confer growth or survival benefits and will occasionally contribute to age-related diseases at the whole-organism level. For example, hematopoietic stem cells carrying *TP53* mutations increase risk for acute myeloid leukemia[9].

To the Editor — Human cell lines are central to biomedical research and medicine, but genetic evolution and inconsistencies among derived lineages are too often ignored. These issues are becoming increasingly important now that wide adoption of gene editing technologies such as CRISPR has led to a boom in the development of new genetic lineages with knock-in reporters or patient-specific mutations (Fig. 1a). A more detailed view of cell line provenance and lineage formation can guard against wasted research effort and funds and, ultimately, improve reproducibility of biological research. Accurate cell line tracking is also required for safely establishing cell therapies for precision medicine.

Currently, 18–36% of common cell lines are estimated to be mislabeled or contaminated; in addition, cell lines often evolve divergent lineages[1,2]. Cell lineages can form by spontaneous or induced selection events during cell culture or when cells are genetically modified. Although funders and journals are starting to acknowledge the importance of cell line authentication, cell lineage provenance is rarely recorded or published, despite its impact on data reliability and reproducibility[3–5].

Here, we discuss lineage divergence as a natural, inevitable phenomenon across all kingdoms of life. We highlight how lineage formation in the culture of human cells is influenced by routine laboratory practices and has accelerated in the genomics and gene-editing era. We also propose simple changes to working routines to minimize unwanted lineage divergence. Lastly, we explore how monitoring divergence can help obtain new biological insights in certain cases.

Lineage formation is ubiquitous in all kingdoms of life

Across life, stochastic genetic changes in clonally proliferating cells lead to de novo lineage formation, affecting both asexually reproducing organisms and somatic cells of sexually reproducing organisms. Long-term evolutionary monitoring of asexual microbe populations, including bacteria and yeast, has yielded critical insights into lineage formation dynamics[6] (Fig. 1b). Lineation can accelerate when stochastic genetic changes result in competitive fitness variation, defined as the time needed for one cell doubling. Although most genetic changes are deleterious or nearly neutral, cells that acquire beneficial genetic changes may outcompete other cells by improved fitness and take over the population. Fitness gains can occur via (i) 'rich-get-richer' effects, in which a beneficial genetic change in a cell drives steady clonal expansion, increasing chances that more changes accrue, or (ii) leapfrog events, in which a genetic change confers an immediate, sizeable fitness benefit to a cell, leading to sudden clonal expansion[7], possibly through emerging epistatic fitness benefits in conjunction with previous changes (Fig. 1c).

Lineage formation also occurs naturally in somatic tissues of multicellular organisms, including plants and animals[8]. It is the driving force of genetic mosaicism and can strongly affect organismal phenotypes. In vivo genetic changes in human tissues result in lineage formation and aging-related mosaicism, such as clonal hematopoiesis. Hematopoietic stem cell lineages accumulate anomalous de novo genetic changes, which can drive detrimental health outcomes (Fig. 1d)[9]. For example, changes in the *JAK2* gene can lead to a 12-fold increase in coronary heart disease risk[9]. Even more dramatic, ~8% of elderly men carry cell lineages without a Y chromosome, which may reduce lifespan by 5.5 years[10]. In cancer, lineage divergence is at the root of complex, genetically heterogeneous tumors. Selection induced by targeted cancer therapies can cause intratumor lineage evolution, with cells escaping therapy regimes, leading to cancer recurrence[11].

When lineage formation occurs in cells or organisms that live in environments under human management, both natural and human-mediated selection will act on phenotypic differences. Charles Darwin divided human-mediated selection into "conscious" and "unconscious" selection. Conscious selection can be an important driving force in human cell culture, for instance, when researchers select cell lineages with increased production of a desired protein[12,13]. However, researchers also unconsciously apply selection pressure on doubling-time differences between cultured cells. Cultured human cells will gradually accumulate differences in this competitive fitness metric, as with microorganisms and somatic cells. Unconscious selection operates continuously and is influenced by researchers' everyday decisions (Fig. 2).

Evolution of human cell lineages during routine cell culture

Evolution drives lineage divergence in all cell types and can strongly affect cellular phenotypes and experimental outcomes. Although cancer cell lines are especially at risk due to their inherent genetic instability[5], similar patterns are observed in other in vitro–grown cell types, such as pluripotent stem cells[14–17].

Perhaps the best-known example of lineage formation and divergence in cultured cells is HeLa[3]. This genetically unstable cervical cancer cell line has been propagated, split and shared among thousands of laboratories since 1951. Analysis of 13 HeLa lineages for genetic and phenotypic differences showed that individual lineages accumulated unique genetic changes after 7–50 passages and chromosome segments varied from one to six copies per lineage[3]. Despite considerable genetic divergence, all independently evolving lineages are referred to in publications as 'HeLa cells'. Only occasionally are the major ancestral lineages, CCL2 or Kyoto, mentioned—and even then, no information on recent pedigree branching or passage number is provided. The phenotypic consequences of lineation include extreme variance in cell doubling time between lineages under identical culture conditions (18–33 h) and discordance in susceptibility to pathobiont infection[3].

Similarly, lineage formation has been observed for the MCF-7 breast cancer and human embryonic kidney 293 (HEK293) cell lines, leading to substantial phenotypic variation between the different cell lineages[4,5], including differences in drug response[5], the ability to grow in suspension[4,12], and gene expression[12].

An analysis of 1,700 lineages from 259 embryonic stem cells (ESCs) or induced pluripotent stem cells (iPSCs) found that ~13% of lineages are aneuploid[15]. Recurrent amplifications of chromosomes 8, 12, 17, 20 and X appear to cause leapfrog events (Fig. 1c), because fitness jumps often occur within ten passages with up to threefold selective growth advantages[14,15,18].

For iPSCs, reprogramming is regularly accompanied by chromosomal amplifications and, overall, iPSCs show similar risks of forming aberrant cell lineages as ESCs[14,15].

Factors influencing cell lineage formation

Several steps and factors in daily laboratory workflows contribute to cell lineage formation and divergence, including cell line establishment, passaging protocols, passage number at the time of an experiment, choice of cell culture media, and freeze–thaw frequency.

The initial steps of cell culture involve a drastic environmental change: cells of all types must adapt to growing in two-dimensional cell culture environments instead of complex three-dimensional tissues in vivo. In these early steps, aneuploidy may confer fitness benefits. A >50% improvement in survival was observed in vitro for ESCs that adapted through specific chromosomal amplifications[17]. Furthermore, the frequency at which a specific genetic alteration occurs in a tumor versus in cell culture can differ[19]. For example, gain of chromosome segment 3q25–27, which is found in 17% of tumor biopsies ($n = 356$), was detected in only 9% of cultured non-small-cell lung cancer cell lines ($n = 86$)[20].

Passaging choices involve fine-tuning of parameters to maintain optimal cell viability and genetic stability[18] (Fig. 2a,b). For stem cells, cell dissociation can be important. For example, enzymatic dissociation disrupts cell–cell contact, which can lead to karyotypic changes[18] or a type of apoptosis called anoikis ('homeless' cell death). Anoikis occurs via cytoskeletal contraction in an isolated cell, activated through Rho-associated kinase (ROCK)-mediated phosphorylation. Although potentially deleterious, full cellular dissociation is required in certain protocols—for example, when establishing monoclonal lineages after genetic engineering. Viable clones derived from single cells experience a genetic bottleneck (Fig. 2c), a risk for any cell type, and thus require characterization. In contrast, full dissociation and subsequent mixing during cell passaging preserves genetic diversity more evenly. The anticipated frequency of anoikis is (unconsciously) factored into decision-making, as dissociative passaging protocols typically recommend 1:1 to 1:5 instead of 1:50 dilutions. An alternative passaging method involves clumped or clustered cell transfer, which maintains cell–cell adhesion. This method minimizes aneuploidy and anoikis[21], and is frequently used for stem cell maintenance. However, with clumped transfer, anomalous genetic changes in transferred cells might be more likely to establish in the descendant cell lineage.

Cell passage number is indicative of the extent of lineage divergence[1,3,5,12,13]. For example, in a model for human intestinal epithelium (Caco-2 cells), lineages with added passages show decreased doubling times, affecting cellular monolayer permeability and drug transport properties[1]. This means between-study variation in experimental results can arise because parallel lineages of the same cell line are compared at different passages. Because these patterns are observed across cell types[1,3,5–7,12,14,15] and may be unavoidable during continuous culturing, cell passage

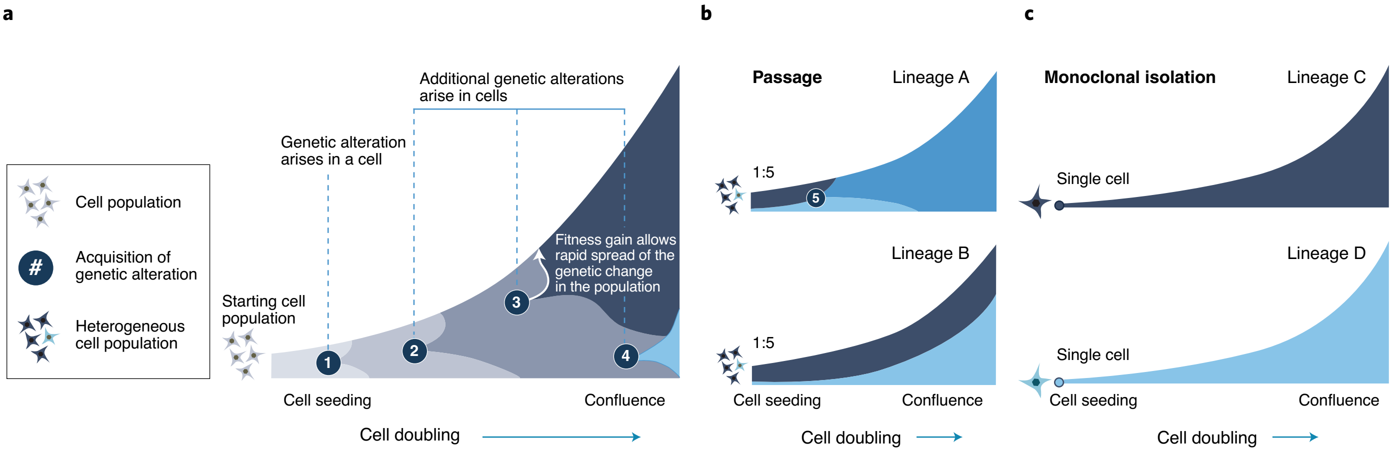


Fig. 2 | **Evolutionary dynamics and selection in cell line establishment and maintenance.** Clonally proliferating cells experience fluctuations in cell density due to population expansion (seeding to confluence) and contraction during subculturing (passaging). In addition, each doubling introduces risk of cells acquiring single-nucleotide or larger changes (circled numbers), such as aneuploidy. Beneficial genetic changes (shortening doubling times) will be selected and sweep through the population, creating divergence between a cell lineage and its ancestors. **a**, The Muller plot shows an example of clonal cell expansion with serial stochastic genetic changes under positive selection resulting in new sublineages. **b**, Passaging of a cell mixture from **a** after confluence (rightmost time point in **a**). The cell population is passaged via dilution (for example, 1:5 split ratio). Assuming perfect mixture, split cell populations have similar levels of sublineage heterogeneity at seeding. When independent de novo genetic changes with fitness benefits continue to accumulate (circled number 5), this results in further lineage divergence into distinct sublineages: observe emergence of the medium blue sublineage in lineage A, creating divergence between lineages A and B. These distinct lineages may have distinct phenotypes. **c**, Isolation of a single cell from **a** at confluence (rightmost time point in **a**). Isolation of single cells for clonal expansion introduces genetic bottlenecks and can result in unintended isolation of sublineages: compare heterogeneous lineages A and B with monoclonal expansion of lineages C and D. For example, the dark gray sublineage in lineage A can thrive as a monoclonal lineage in lineage C, but would otherwise have disappeared.

number and doubling speed should be recorded so researchers can compare between experiments.

Choice of media can also alter adaptive characteristics of cell survival and growth. ROCK inhibition reduces cytoskeletal contraction during dissociative passaging and promotes cell survival[16]. Inhibitors are therefore widely used during dissociation and single-cell cloning steps of ESC and iPSC culture protocols. However, because cells with highly deleterious genetic changes can be removed through apoptosis, blocking this with ROCK inhibitors may increase the risk of selecting novel lineages with aberrant genetic changes. Medium regimens are further used to direct cell adaptation for specific characteristics. For example, the HEK293F lineage, cultured for factor VIII production to treat hemophilia[13], was adapted to serum-free suspension culture from a fast-growing adherent clone. This consciously selected lineage is now considered stable for this phenotype and can be distinguished from related lineages through gene expression profiling[12].

Freezing cell populations at specific passages allows researchers to halt ongoing lineage formation and assay specific passages along a lineage (Fig. 1a). We have shown the viability of this approach by performing high-throughput sequencing on a DNA-barcoded mouse cancer cell line before and after freeze–thaw, finding cell barcode distributions were preserved after recovery ($r^2 \approx 0.9$)[22]. However, caution is warranted, as freeze–thaw cycles can still pose a genetic bottleneck when working with some cell lines or lineages. Indeed, the genetic copy number profile of HEK293S-lineage cells shifted over freeze–thaw cycles[4]. Cells that survive freeze–thaw and outcompete others determine the genetic makeup of the thawed population, leading to changes in doubling time and other traits. Doubling times of recently thawed cells may continue to change during additional passaging[4,5].

This is a subset of daily laboratory routines influencing in vitro lineage formation and divergence. Researchers should take measures to monitor and minimize lineage divergence in cell lines during routine culturing (Table 1), which is important given the recent increase in derivation of new cell lines and lineages.

Accelerated lineage formation in the genomics and gene-editing era

Over the past decade, new genome engineering technologies have made it straightforward to knock out or tag genes and insert putative disease variants[21]. Given the rapid adoption of genome editing tools, the boom in new cell lineages is poised to continue (Fig. 1a). New lineages can be derived by engineering genetic changes using programmable nucleases (for example, zinc fingers, TALENs and CRISPR–Cas proteins), integrating transposons, viral vectors or plasmids[4].

CRISPR-mediated gene disruption and homology-directed repair now make it straightforward to edit most human cell lines[23,24]. Although CRISPR nucleases have accelerated the pace of genome editing, they can also induce off-target genetic changes[25], which necessitates careful monitoring by whole-genome sequencing, targeted sequencing of predicted off-targets and/or traditional karyotyping (Table 1)[26]. CRISPR editing in stem cells induces a p53-dependent DNA damage response and, through unconscious selection, can enrich for surviving cells carrying oncogenic p53 mutations[27,28].

After genome engineering in iPSCs, derivation of monoclonal lineages can take ~14 passages, spanning approximately two months[24]. Such an essential cloning step poses an extreme genetic bottleneck on cell populations. Cells with unintended de novo genetic changes might be selected as founders of new monoclonal lineages, posing major risks for ongoing experimentation (Fig. 2b,c)[25,26]. To protect against these, multiple engineered cells

Table 1 | Roadmap for routine cell tracking

Step	Action	Methods
Start	Create a journal devoted to cell lines used in your laboratory.	Create a designated cell-tracking workbook using Google Sheets. One of us (S.Z.) has started a company (FIND Genomics, https://findgen.bio) to develop specialized cell-tracking software that integrates laboratory record-keeping with genome sequencing using low-pass whole-genome sequencing (WGS).
Track	Record daily routine steps.	Write down time between passages, total days in vitro, splits, use of specific media, treatments, protocols or modifications, as well as observations on cell doubling time and cell health.
	Track individual cell lineages.	When a cell population is passaged (split) from a single to multiple culture flasks, track ongoing passaging separately for each individual new lineage.
	Record which cell population was used in an experiment.	Mark usage of a specific cell population in an experiment. Indicate which cell lineage the cell population is part of.
	Assess lineages for the presence of *Mycoplasma* infection.	Record which cell populations are tested for *Mycoplasma* within each lineage. This allows tracing back which cell lineages are at risk and which ones can be salvaged.
Genotype	Periodically authenticate cell populations from specific lineages.	Assess short tandem repeat (STR) or single nucleotide polymorphism (SNP) markers offered by service providers. Alternatively, use low-pass WGS to match your cell population to your cell line database. Depending on experimentation intensity, this should ideally be done once every 15 passages. High-risk cell lines are listed by the International Cell Line Authentication Committee.
	Periodically determine genetic stability of cell lineages. Especially needed after creating novel monoclonal lineages.	Analysis of genetic stability of cell populations can be meaningful at various levels of resolution: karyotype-, copy number-, mutation- or site-specific changes. The resolution required will depend on cell type and experimental goals. The chosen resolution will determine the cost of data collection per sample using low-pass WGS or other assays. Record the passage number and the resolution of the test performed at the time of testing.
	Look for off-target effects from CRISPR engineering.	Use WGS or specialized methods like break labeling in situ and sequencing (BLISS) or circularization for in vitro reporting of cleavage effects by sequencing (CIRCLE-seq).
Share	Record all exchanges of cell lines with other laboratories. When intending to repeat an experiment from a particular laboratory, ask this laboratory for cells from the exact same lineage used in the published experiment(s). In this manner, discrepancies between experiments can be traced back to a particular lineage or passage number.	Ask the providing laboratory to share a cell lineage's passage number and history (for example, information from the Track and Genotype steps above).
	Verify lineage evolution to ensure possible experimental discrepancies are not due to lineage divergence.	Compare cell doubling time and genetic stability for cells currently in culture with the original cell population from the originating laboratory.
Search and update	Register any new cell lines or genetically distinct cell lineages (formed through conscious or unconscious selection).	Examples of public cell databases are hPSCreg or Cellosaurus.

should be characterized genetically and phenotypically to ensure that at least several clonal lineages are representative of the original cell line[29]. This is a good practice for both engineered and non-engineered cell cloning.

Subsequently, established cell lineages should be named and can even be published as stand-alone resources in scientific journals. For example, we have recently developed a pluripotent human ESC line (HUES66) with an inducible lentiviral construct to trigger cortical neuron differentiation[24]. After verifying pluripotency and normal karyotype, we named the new CRISPR-engineered cell lineage and registered it in a database (hPSCreg)[30].

Despite the advantages of registering cell lines in searchable databases, such as the ability to search for particular cell lines or lineages quickly, currently this is not standard practice. It will be important to understand impediments that prevent wider use of cell line databases. A repository for cell line and lineage sharing would be transformative as the number of engineered cell lineages continues to grow rapidly.

In summary, there is an urgent need to monitor cell line genetic identity and stability routinely. In the long term, lineage tracking can increase research efficiency, enhance the breadth of insights gleaned from experiments, and improve the ease of meeting publication requirements for documenting the cell lines and methods used. Most importantly, it can lead to more reproducible science (Table 1). Above all, documenting daily cell line management routines and being cognizant of cell line provenance and lineage formation is an important step toward improving experiment reproducibility and interpretability.

Sophie Zaaijer[1,2,5] ✉, Simon C. Groen[3,5] ✉ and Neville E. Sanjana[3,4] ✉

[1]Cornell Tech, New York, NY, USA. [2]FIND Genomics, New York, NY, USA. [3]Department of Biology, New York University, New York, NY, USA. [4]New York Genome Center, New York, NY, USA. [5]These authors contributed equally: Sophie Zaaijer, Simon C. Groen.

✉*e-mail: sophie@findgen.bio; sc.groen@nyu.edu; neville@sanjanalab.org*

Published online: 19 May 2021

https://doi.org/10.1038/s41587-021-00928-1

References

1. Hughes, P., Marshall, D., Reid, Y., Parkes, H. & Gelber, C. *Biotechniques* **43**, 575, 577–578, 581–582 (2007).
2. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. *PLoS Biol.* **13**, e1002165 (2015).
3. Liu, Y. et al. *Nat. Biotechnol.* **37**, 314–322 (2019).
4. Lin, Y.-C. et al. *Nat. Commun.* **5**, 4767 (2014).
5. Ben-David, U. et al. *Nature* **560**, 325–330 (2018).
6. Boyer, S., Hérissant, L. & Sherlock, G. *PLoS Genet.* **17**, e1009314 (2021).
7. Nguyen, Ba,A. N. et al. *Nature* **575**, 494–499 (2019).
8. Myles, S. et al. *Proc. Natl. Acad. Sci. USA* **108**, 3530–3535 (2011).
9. Watson, C. J. et al. *Science* **367**, 1449–1454 (2020).
10. Forsberg, L. A. et al. *Nat. Genet.* **46**, 624–628 (2014).
11. Fittall, M. W. & Van Loo, P. *Genome Med.* **11**, 20 (2019).
12. Malm, M. et al. *Sci. Rep.* **10**, 18996 (2020).
13. Dumont, J., Euwart, D., Mei, B., Estes, S. & Kshirsagar, R. *Crit. Rev. Biotechnol.* **36**, 1110–1122 (2016).
14. Lund, R. J., Närvä, E. & Lahesmaa, R. *Nat. Rev. Genet.* **13**, 732–744 (2012).
15. Taapken, S. M. et al. *Nat. Biotechnol.* **29**, 313–314 (2011).
16. Weissbein, U. et al. *iScience* **11**, 398–408 (2019).
17. Barbaric, I. et al. *Stem Cell Reports* **3**, 142–155 (2014).
18. Bai, Q. et al. *Stem Cells Dev.* **24**, 653–662 (2015).
19. Sharma, S. V., Haber, D. A. & Settleman, J. *Nat. Rev. Cancer* **10**, 241–253 (2010).
20. Yamamoto, H. et al. *Cancer Res.* **68**, 6913–6921 (2008).
21. Komor, A. C., Badran, A. H. & Liu, D. R. *Cell* **168**, 20–36 (2017).
22. Chen, S. et al. *Cell* **160**, 1246–1260 (2015).
23. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. *Nat. Biotechnol.* **34**, 339–344 (2016).
24. Lu, C. & Sanjana, N. E. *Stem Cell Res.* **41**, 101643 (2019).
25. Kosicki, M., Tomberg, K. & Bradley, A. *Nat. Biotechnol.* **36**, 765–771 (2018).
26. Przewrocka, J., Rowan, A., Rosenthal, R., Kanu, N. & Swanton, C. *Ann. Oncol.* **31**, 1270–1273 (2020).
27. Ihry, R. J. et al. *Nat. Med.* **24**, 939–946 (2018).
28. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. *Nat. Med.* **24**, 927–930 (2018).
29. Kimberland, M. L. et al. *J. Biotechnol.* **284**, 91–101 (2018).
30. Isasi, R., Namorado, J., Mah, N., Bultjer, N. & Kurtz, A. *Stem Cell Res.* **40**, 101539 (2019).

Acknowledgements

We thank L.M. de Buy Wenniger and T. Joseph for valuable feedback on the manuscript and K. White for assistance with graphic design. S.Z. is supported by the Jacobs Technion-Cornell Institute, the Elisha M. Friedman Postdoctoral Fellowship and FIND Genomics. S.C.G. is supported by a fellowship from the Gordon and Betty Moore Foundation/Life Sciences Research Foundation through grant GBMF2550.06. N.E.S. is supported by the National Institutes of Health (NIH) National Human Genome Research Institute (grant no. DP2HG010099), NIH National Cancer Institute (grant no. R01CA218668), NIH National Institute of General Medical Sciences (grant no. R01GM138635), Defense Advanced Research Projects Agency (grant no. D18AP00053), Sidney Kimmel Foundation, Melanoma Research Alliance, Brain and Behavior Foundation and Cancer Research Institute, and New York University and New York Genome Center startup funds.

Author contributions

Conception: S.Z., S.C.G., N.E.S. Writing and figures: S.Z., S.C.G., N.E.S.

Competing interests

S.Z. is a co-founder of FIND Genomics, a company that aims to improve reproducible cell-based science and develops cell lineage tracking software. N.E.S is an advisor to Vertex.

Additional information

Peer review information *Nature Biotechnology* thanks Chad Cowan and Johan Rockberg for their contribution to the peer review of this work.