An engineering approach to decode immune responses

Bresser, K.
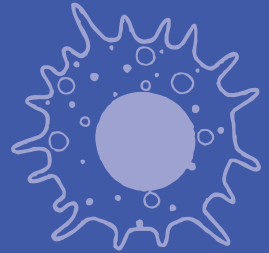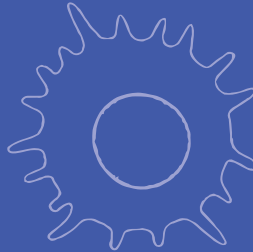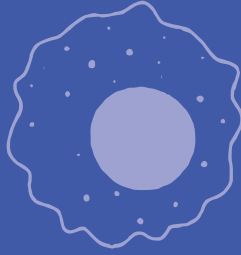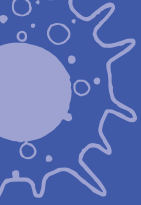
Chapter 6.
Discussion

OR: "Lets hear some answers to
questions that nobody asked."
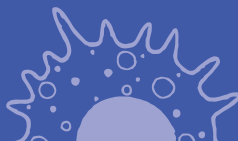
While making my first steps into the academic world I encountered a host of tenacious dogmas, enlightening opinions, and close-to-ideal-but-not-so-perfect systems. Although this journey has caused me to consider, and re-consider, many different topics included in this thesis, I will use this this chapter to dive a bit deeper into two—perhaps somewhat unrelated—themes that have fascinated me during my work. First, I will consider the current theories on stemness in the CD8$^+$ memory T cell population and offer my own view on this topic, a view that was strongly shaped by our work in chapter 4. Second, I will discuss the current structure for academic data-sharing that should allow for the re-use (or re-purposing) of data and findings. In chapter 3 through 5 I sought to use data from others to enrich our own observations, but found that this was not a simple feat, in part forming the basis for my interest in the matter. In the final section of this part of the discussion, I will give an example of some of the current problems with the accessibility of scientific data, and opine on a number of ways to improve the manner in which data is shared across the academic community.

Stemness in the CD8$^+$ memory T cell pool

My colleagues and I ended the discussion of **chapter 4** with the notion that the organizing principles of the T_{CM} pool shares some similarities with the stem cell compartments found in solid tissues. Stemness in the CD8$^+$ memory T cell pool has been widely discussed in the literature, with the concept of stemness taking many shapes and forms. Although the premise of true stem cell activity is, at first glance, somewhat counterintuitive for a highly differentiated cell type such as T cells, I do feel that this concept provides a helpful framework when considering lineage relationships in the T cell pool.

In the following sections I will outline some influential studies investigating stemness in the CD8$^+$ T cell memory pool, describe where, in my view, these studies have placed our understanding of this concept, and additionally suggest a few topics that I believe require further consideration.

Stemness in the context of T cell immunity

Stemness is commonly defined as the capacity of a cell pool to allow both *self-renewal* (duplicating oneself in relative perpetuity) and *differentiation* (regenerating a functional tissue)[1]. In many anatomical compartments, this property is restricted to a small subset of multipotent cells capable of differentiating into the specialized cells that make up the tissue[2–4]; a process that is accompanied with the progressive loss of stemness. This concept also holds true for the majority of immune cells, such as monocytes and neutrophils, which are continuously replenished by hematopoietic stem cells (HSCs) in the bone marrow. However, due to their adaptive nature, T and B cells cannot rely on this replenishment model. Specifically, the naïve T cell pool comprises an immense variety of T cell receptor (TCR) clonotypes, generated though random re-arrangement of gene-fragments. Upon pathogen encounter, relevant antigen-specific T cell clonotypes are selected to expand, differentiate and combat the ongoing infection, and subsequently establish a long-lived memory pool. This memory T cell pool retains the capacity to repeatedly differentiate and expand upon multiple cycles of infection, but must do so independent of de novo generation from the bone marrow, in order to retain the critical clonotype information. This highlights an interesting question in the developmental hierarchy of T cell memory; *How to allow for successive rounds of proliferation and differentiation without reinforcements from HSCs?*

An attractive hypothesis for the maintenance of T cell memory is that, analogous to other tissues, stemness or multipotency is restricted to a minute subset. Evidence in favor of this model came from a series of studies describing a small subset of memory T cells that existed in a multipotent naïve-like state, and this population has been coined 'memory stem cells' (or T_{SCM})[5–7]. This T_{SCM} population possessed superior proliferative potential and retained a high level of multipotency upon TCR stimulation. Furthermore, the T_{SCM} pool was phenotypically similar to the T_{CM} pool apart from the peculiar retention of CD45RA, a protein that has been extensively used by immunologists as a mark for naïve T cells. Transcriptional profiling put these antigen-experienced cells at the apex of the memory T cell pool hierarchy[7,8], placing them somewhere between naïve and central memory T cells.

Although the presence of T_{SCM} provides a parsimonious model for developmental hierarchy in the memory T cell pool, this model does not fully fit with adoptive cell transfer studies showing that the ability to give rise to secondary effector pools is abundantly present in the T_{CM} pool[9,10]. This stem-like capacity of T_{CM} was elegantly demonstrated by Graef *et al.*[11], showing that single CD62L[+] T_{CM} cells (that had not been specifically selected for the expression of T_{SCM} associated markers) were able to reconstitute a functional T cell pool throughout multiple successive rounds of single cell transfer. This study thereby established that stemness is a characteristic shared across many cells in the T_{CM} pool.

So how does this finding fit with the T_{SCM} model? A simple consideration provides some insight here: As demonstrated in **chapter 4**, the T_{CM} pool comprises cells that exhibit a variety of transcriptional states and distinct behaviors. Therefore, the transcriptome analyses of bulk T_{SCM} and T_{CM} that placed T_{SCM} as a more naïve-like subset relative to T_{CM}, may have been confounded by the latter's internal heterogeneity. In line with this possibility, a recent study by Galletti *et al.* showed that the depletion of PD1[+]TIGIT[+] cells from the T_{SCM} and T_{CM} pools largely eliminated the transcriptional and functional differences between these two cell populations[12]. Furthermore, findings from a scRNAseq study of tumor-infiltrating and blood-derived T cells suggest that there is a noteworthy degree of promiscuity in the expression of CD45 isoforms across T cell subsets[13]. These findings could imply that T_{SCM} should not be considered an entirely distinct population as previously imagined, but rather a constituent of the T_{CM} pool that is primarily set apart by its alternative splicing of CD45. Whether the differential expression of distinct splice-forms of CD45 is functionally relevant in memory maintenance or re-expansion potential will be an interesting topic for future endeavors.

In summary, the studies by Galletti and Graef would support a model in which the T_{CM} pool, as a whole, serves as a stem-like reservoir maintaining each T cell clone. While stemness appears to be a shared property of T_{CM}, our findings, in addition to those made by Galletti and colleagues, signify that a degree of specialization is present within this population. This begs the question: If all T_{CM} are equal, are some T_{CM} more equal than others?

Decreasing potential or division of labor

The phenotypic and transcriptional diversity within the T_{CM} population that we observe in **chapter 4** presents us with two intriguing avenues to pursue. First, it will be interesting to explore how this diversity is established. Our observation that prior cell division is correlated with an effector-like transcriptional program is consistent with several models. For instance, this could mean that at some point during the acute phase of the T cell response T_{CM} precursors diverge, with one lineage "deactivating" into a more naïve-like quiescent behavior, whereas the other lineage maintains a more activated state and continues to divide. Alternatively, effector-like T_{CM} cells could derive from a separate lineage of de-differentiated effector T cells[14,15], which could explain their transcriptional state and extensive degree of prior proliferation. A strategy to experimentally address this issue would be to leverage the *Klrg1^{Cre}* mice developed by Herndler-Brandstetter *et al.*[14], to examine whether the effector-like T_{CM} pool is enriched for cells that have previously expressed the effector T

cell-associated protein KLRG1.

A second valuable direction will be to investigate the functional relevance of T_{CM} diversity upon re-infection. In our efforts we have found substantial differences in transcriptional profiles and re-expansion potential within the T_{CM} pool. This fits with an observation made in the Galletti study[12] that a more differentiated T_{CM} sub-population, that the authors term "pre-exhausted", exhibits reduced replicative potential. A matter that is currently unresolved is whether this more differentiated state indicates that these cells are of little value or, rather, that the diversity in cell states in the T_{CM} pool is reflective of a division of labor. Some experiments from **chapter 4** and Galletti *et al.* may provide some insight here. Specifically, our effector-like T_{CM} appear to degranulate to a larger extent upon short term ex vivo stimulation. Likewise, Galletti's "pre-exhausted" T_{CM} contain more accessible chromatin at cytotoxicity-related genomic loci. These observations could indicate a degree of specialization within the T_{CM} pool, in which some cells are more prone to re-expand and others are predisposed to rapidly re-exert effector functions.
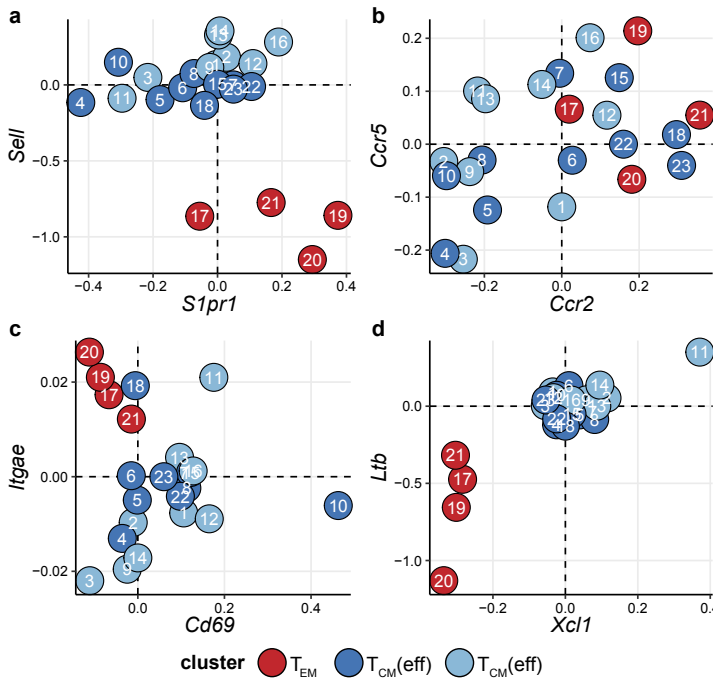
Figure 1. T_{CM} exhibit heterogenous mRNA levels of various genes involved in cell trafficking and localization. Differential expression of selected genes across the T_{CM} and T_{EM} MetaCells identified in **chapter 4**. (**a-d**) Log2 transformed gene-expression enrichment is plotted for [**a**] *Sell* (CD62L) versus *S1pr1* (Sphingosine-1-phosphate receptor 1), [**b**] *Ccr5* versus *Ccr2*, [**c**] *Itgae* (CD103) versus *Cd69*, and [**d**] *Ltb* (Lymphotoxin beta) versus *Xcl1* (Lymphotactin)

Nature versus nurture at the cellular level

The studies discussed above provide compelling evidence that the T_{CM} pool comprises cells that exist in distinct states, and that differ in their capacity to execute specific functions upon re-activation. But what underlies this disparity during a recall response? Is, for instance, the enhanced re-expansion potential of quiescent T_{CM} fully attributable to their cell state (i.e., their *nature*), or could the biased localization of different T_{CM} types in distinct niches (i.e., *nurture*) play a role?

The lymphoid tissues in which T_{CM} largely reside (such as lymph nodes and spleen) have a complex organization of myeloid, lymphoid, and stromal cells, compartmentalizing these organs into distinct niches. Several studies have highlighted the importance of memory T cell positioning within secondary lymphoid organs to recall responses, and demonstrated a key role for chemokine receptors in this process[16–18]. However, such studies have generally not assessed whether T_{CM} with different cell states are differentially positioned. Re-examination of the scRNAseq dataset of splenic CD8+ memory T cells presented in **chapter 4** offers some clues on this matter. While all T_{CM} expressed high levels of *Sell* transcripts (encoding the lymphoid-tissue entry receptor CD62L), these cells displayed heterogeneous expression of the tissue-egress associated gene *S1pr1* and several chemokine receptors (**Fig. 1a-b**). Several T_{CM} MetaCells additionally differed in their expression of *Cd69* and *Itgae* (encoding CD103), genes classically associated with tissue-resident memory T cells (**Fig. 1c**). Interestingly, one of these MetaCells with elevated *Cd69* and *Itgae* expression was additionally marked by relatively high levels of *Ltb* and *Xcl1* transcripts (**Fig. 1d**). Both of these genes encode secreted factors for which the receptors are present on the myeloid and stromal component of lymphoid tissues[19,20]. The differential expression of these cell migration and retention-associated genes could therefore imply that these T_{CM} subtypes possess a different affinity toward specific local niches.

As a final note, if such differential positioning of T_{CM} indeed underlies distinct functional outcomes, it would be highly interesting to investigate the stability of these niches. Specifically, are these niches seeded upon memory formation and subsequently remain immutable, or is there a certain degree of plasticity, allowing T_{CM} to move in and out of these niches? Also, can such niches exclusively be seeded during resolution of infection, or is simply the correct expression of specific chemotactic receptors/factors enough? This latter question may be particularly noteworthy, as its answer would strongly affect the manner in which studies using the re-transfer of T cell subsets (that are taken out of their original niche) should be interpreted. By the same token, the mechanism of niche formation could have implications for T cell based cellular therapies, as putative factors necessary for the establishment of niches that ensure long-lasting protective T cell responses may not be sufficiently present. In the event that improper niche formation negatively impacts T cell immunity, such cellular therapies may conceivably be modified to incorporate this component, for instance through the use of adjuvants (e.g., cytokines or chemokines) or forced expression of putative niche-inducing factors through genetic engineering of transferred T cells.

Scientific equity through data sharing

Scientific discovery is, at its core, a community effort, with each new insight being built on the foundation of data that was provided by predecessors. Therefore, I feel it is important that our precious data is viewed not only as a means to an end for our specific question, but first and foremost as a starting point for others. This would entail properly storing and sharing our published findings in a way that is easily accessible and interpretable by others. In **chapters 3, 4 and 5** I sought to validate specific findings using data from others, and found that the habit of proper data storage is still far from commonplace in the scientific community. Although I could usually find plenty of studies that contained experiments useful to my research questions, a large share of this data was either not published alongside the article at all or uploaded in a manner that did not allow for proper re-use.

By no means am I suggesting that all is Fire and Brimstone, in fact, data sharing is a field of lively discussion and steady (albeit slow) innovation. In the next few paragraphs, I will briefly outline the most prominent philosophy for data sharing, then provide an example of a data type for which we are still 'playing catch-up', and finally discuss my view on the role of three major stakeholders/ components of the scientific community (publishers, repositories, and scientists) in the improvement of data availability.

A philosophy for open science

In practice, the academic field is purposefully unequal, resulting in disparities in the ability of different labs to generate certain types of datasets. Because the source of this inequality is difficult to address (e.g., there cannot and should not be infinite funding for all labs), data sharing provides a way toward more *equitable outcomes*, as everyone would be able to reap the benefits of data obtained by a few. To achieve such an outcome, a number of parties from various disciplines met in Leiden (2014) to discuss the principles of open science. During this workshop, the FAIR principles[21] were drafted, which represent a general philosophy of data sharing that can be applied broadly in the scientific community. Essentially, for a dataset to be FAIR it needs to be *Findable*, *Accessible*, *Interoperable* and *Re-usable*, factors that are mostly determined by the richness of information on experimental conditions or outcomes of analyses that is shared alongside the data, referred to collectively as metadata.

Say I have found an intriguing gene-expression network in my pet cell type, and now would like to pressure-test these findings in an external RNAseq dataset. First, I should be able to *find* a relevant dataset, meaning the original research article should refer me to the RNAseq data through a permanent link or should be easily found through a query of the relevant repository. In the latter case, rich metadata detailing the experiment is key, as it simplifies discovery through a search engine. Second, I should be able to *access* and download the data freely without the need of going through a paywall or creating some site-specific account, to the extent that privacy regulations permit. Third, the dataset needs to be interpretable (both to machines and humans) so I can integrate it in any analysis—through its *interoperability*—meaning that the files use standardized formats and its

annotations use vocabularies that are widely applied (e.g., genes are identified with gene symbols or ensemble identifiers). Last, I should be able to easily use, or rather *re-use*, the obtained data for its new purpose. To achieve this, rich metadata is again crucial, providing machine-readable sample-level information on both experimental conditions and downstream analyses (e.g., outcomes of a clustering analysis or the code that was applied to generate the manuscript figures). If the data produced by a study is shared in a FAIR manner, a secondary user should be able to use and mine the dataset within a matter of hours, providing the user with a means to start asking questions the original authors have not considered.

## The case of single cell RNA sequencing

Single cell sequencing is a good example where FAIR data sharing is highly beneficial. Due to their richness of information, a multitude of questions can be probed in each dataset, making these datasets useful reference points for research lines outside of their original intended purpose. In order to serve as such reference points, it is key that access is provided to both the raw data *and* the results of all down-stream analyses. This means that detailed metadata—*at the cell level*—is crucial, as it allows direct re-use and integration of a study's results for new endeavors.

Unfortunately, assessment of single cell sequencing datasets deposited to the Gene Expression Omnibus (GEO) over the last couple of years shows that the majority of datasets does not include identifiable metadata (**Fig. 1a**). While such data can still be re-analyzed from scratch, a direct comparison to the authors results is severely complicated. The observed inconsistency in the FAIRness of single cell sequencing datasets uploaded to GEO is not entirely surprising, as the information page '*Submitting high-throughput sequence data to GEO*' does not provide information or guidelines on submitting single cell sequencing data (as of this writing). Furthermore, a lot of variability can be found in the data availability guidelines among the different scientific publishers. Some publishers, such as Cell Press[22], leave little ambiguity, whereas many other publishers mainly provide either vague or dated guidelines[23,24]. Interestingly however, these differences do not appear to result in better or worse commitment to data re-usability, as the percentage of depositions that include metadata are comparable between the different publishers (**Fig. 1b, c**).

As a final note, improved FAIRness of scRNAseq data appears to be positively associated with the influence of a manuscript. Specifically, manuscripts that included metadata in their data depositions are generally cited to a higher degree as compared to those that do not (**Fig. 1d**). This trend could potentially indicate that proper data deposition increases the likelihood that others will re-use the data and thus reference the original manuscript in their work.

Where can we improve things?

Publishers. As gatekeepers of peer-reviewed scientific content, academic journals play a pivotal role in FAIR data sharing. As mentioned above, many publisher guidelines on data sharing are written in an *implicit* manner, suggesting various repositories and requesting adherence to community
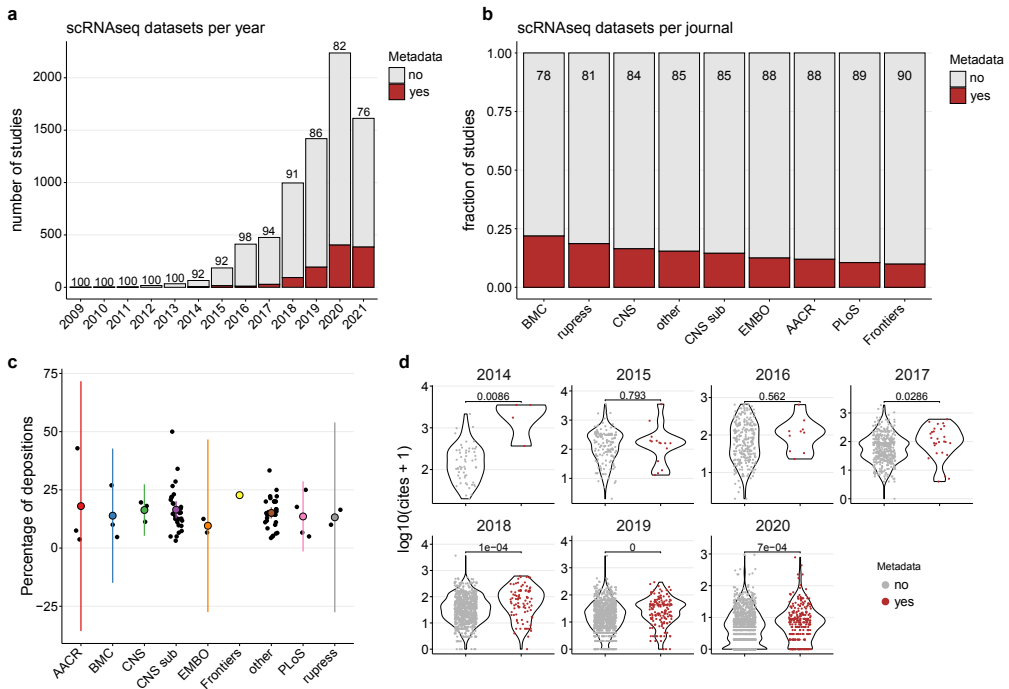
Figure 2. Inclusion of metadata in scRNAseq datasets deposited to GEO. (a) Number of depositions that did or did not include metadata in each year since 2009. Numbers on top of stacks denote the percentage of depositions without metadata. **(b)** Fraction of depositions that did or did not include metadata per publisher. Numbers on top of stacks denote the percentage of depositions without metadata. **(c)** Percentage of depositions that did or did not include metadata per publisher. Black dots indicate individual journals, colored dots indicate means, colored lines represent the 95% confidence interval. **(d)** Number of citations that a manuscript received since publication. Depicted as violin plots, dots indicate individual manuscripts. *P* values indicated in the plots were calculated by Wilcoxon signed-rank test, followed by Holm-Bonferroni correction.

standards such as the FAIR principles. I feel a more effective approach would be to make such guidelines *explicit*. For example, journals could provide a mock-up manuscript in which various commonly used data types are used, showing an impeccable example of how the different data components can be deposited and shared. The same mock-up would ideally be used across multiple journals from the same or different publishers to achieve homogeneous standards.

Peer-review offers another opportunity to ensure that the data underlying the results of the study can be readily assessed. The peer-review system is in place to ensure that published content is valuable to the community, it therefore makes sense to allocate more weight to data availability in the assessment of a manuscript. This could be achieved by either requesting reviewers to include an analysis of the efforts that were made by the authors to adhere to the FAIR principles in their assessment, or appointing a specialized reviewer whom specifically covers this aspect. To simplify this process, a short checklist could be offered to reviewers specifying points of interest. For instance, if both raw and processed data can be found and downloaded easily, and whether field-relevant repositories are used.

Repositories. As data repositories define how results are deposited, these entities are in a crucial position in the data sharing network. In my view, a big leap in the right direction would be for repositories to harmonize their guidelines with the scientific publishers. This would, as noted in the segment above, include the specification of explicit instructions on the contents of a deposition that match the requirements of major publishers and dummy uploads that are linked to mock-up manuscripts.

Furthermore, setting strict requirements for the inclusion of metadata and processed data alongside raw data would be appropriate. This prevents repositories from turning into 'data dumpsites', where the findings are *technically* shared, but re-use is severely complicated. Using scRNAseq uploads as an example, this would entail that sequencing results should be supplemented with *at least* one metadata file (e.g., results from clustering or pseudotime analyses) and *at least* one processed data file (e.g., results from gene-set enrichment or custom analyses). Again, implementation of these requirements will work best if they are set in collaboration with the scientific publishers.

Scientists. In my view, both publishers and repositories have a huge influence in shaping the data sharing environment. However, I do not believe that a perfect system can be built if we would have these two bodies policing the FAIRness of all data uploads. They should be here to guide and enable the process, but the ultimate responsibility needs to lie with the ones generating the data, the scientists.

Improving the way that scientists treat their data is not trivial; to be efficacious—at least in the long term—FAIR practices need to be instilled into the culture of the community. This would begin at the university level, for example including primers in the curriculum that discuss best practices on documenting one's findings. Next, internships provide a perfect microcosm for an applied scientific project, making them crucial moments in teaching scientists-to-be the importance of reporting re-usable data. In practice, the supervisor could provide the student with a system that would make the obtained data FAIR within the lab, and essentially ready to publish. It would also be desirable to integrate the importance of finishing the internship with re-usable data in the grading system, giving this aspect equal weight as, for instance, quality of the practical work.

In large part, the lab culture is defined by its principal investigator. Therefore, it is important for them to provide guidance in, and promote adherence to, a FAIR system of data storage. This would not necessarily require group leaders to micro-manage filing systems used by their scientists, but can simply entail being an advocate for the desired ideals. For example, engaging in discussions on open science during meetings or encouraging diligence in these topics during the final stages of a project.

Concluding remarks

Everything taken together, it will take a concerted effort from all members of the scientific community to shape and—most importantly—maintain a FAIR data sharing network. As I have noted, I feel that a tight collaboration between publishers and repositories will result in a huge step

forward. Such a unified front would eliminate much of the ambiguity that currently complicates data deposition. Furthermore, it will likely be key to provide both scientists and publishers with appropriate incentives to implement the FAIR principles. For example, by more prominently integrating adherence to these principles into journal impact metrics, or awarding/penalizing investigators during the allocation of publicly funded research grants.

Technologies and their resulting datasets will continue to evolve, meaning that keeping FAIRness at a constant level represents a Red Queen's race. I feel it is important that we, as a community, do not lose sight of this principle, and remain flexible in the implementation of changes in order to 'keep up'.

References

1. Laplane, L. Cancer stem cells: philosophy and therapies. (Harvard University Press, 2016).
2. Ito, M. et al. Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. Nat. Med. 11, 1351–1354 (2005).
3. Yan, K. S. et al. The intestinal stem cell markers Bmi1 and Lgr5 identify two functionally distinct populations. Proc. Natl. Acad. Sci. U.S.A. 109, 466–471 (2012).
4. Wilson, A. et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. Cell 135, 1118–1129 (2008).
5. Zhang, Y., Joe, G., Hexner, E., Zhu, J. & Emerson, S. G. Host-reactive CD8+ memory stem cells in graft-versus-host disease. Nat Med 11, 1299–1305 (2005).
6. Gattinoni, L. et al. Wnt signaling arrests effector T cell differentiation and generates CD8+ memory stem cells. Nat. Med. 15, 808–813 (2009).
7. Gattinoni, L. et al. A human memory T cell subset with stem cell-like properties. Nat. Med. 17, 1290–1297 (2011).
8. Fuertes Marraco, S. A. et al. Long-lasting stem cell–like memory CD8 + T cells with a naïve-like profile upon yellow fever vaccination. Sci. Transl. Med. 7, (2015).
9. Bouneaud, C., Garcia, Z., Kourilsky, P. & Pannetier, C. Lineage relationships, homeostasis, and recall capacities of central- and effector-memory CD8 T cells in vivo. J Exp Med 201, 579–590 (2005).
10. Wherry, E. J. et al. Lineage relationship and protective immunity of memory CD8 T cell subsets. Nat Immunol 4, 225–234 (2003).
11. Graef, P. et al. Serial transfer of single-cell-derived immunocompetence reveals stemness of CD8(+) central memory T cells. Immunity 41, 116–126 (2014).
12. Galletti, G. et al. Two subsets of stem-like CD8+ memory T cell progenitors with distinct fate commitments in humans. Nat Immunol 21, 1552–1562 (2020).
13. Zhou, B. et al. Single-cell RNA-Seq reveals the link between CD45 isoforms and tumor-infiltrating T cells heterogeneity in liver cancer. http://biorxiv.org/lookup/doi/10.1101/2020.03.22.002824 (2020) doi:10.1101/2020.03.22.002824.
14. Herndler-Brandstetter, D. et al. KLRG1+ Effector CD8+ T Cells Lose KLRG1, Differentiate into All Memory T Cell Lineages, and Convey Enhanced Protective Immunity. Immunity 48, 716-729.e8 (2018).
15. Youngblood, B. et al. Effector CD8 T cells dedifferentiate into long-lived memory cells. Nature 552, 404–409 (2017).
16. Kastenmüller, W. et al. Peripheral Prepositioning and Local CXCL9 Chemokine-Mediated Guidance Orchestrate Rapid Memory CD8+ T Cell Responses in the Lymph Node. Immunity 38, 502–513 (2013).
17. Böttcher, J. P. et al. Functional classification of memory CD8+ T cells by CX3CR1 expression. Nat Commun 6, 8306 (2015).
18. Bangs, D. J. et al. CXCR3 regulates stem and proliferative CD8+ T cells during chronic infection by promoting interactions with DCs in splenic bridging channels. Cell Reports 38, 110266 (2022).
19. Shou, Y. et al. Redefining the Role of Lymphotoxin Beta Receptor in the Maintenance of Lymphoid Organs and Immune Cell Homeostasis in Adulthood. Front. Immunol. 12, 712632 (2021).
20. Crozat, K. et al. The XC chemokine receptor 1 is a conserved selective marker of mammalian cells homologous to mouse CD8α+ dendritic cells. Journal of Experimental Medicine 207, 1283–1292 (2010).
21. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
22. Information for Authors: Cell. https://www.cell.com/cell/authors.
23. Supplementary information | Nature. https://www.nature.com/nature/for-authors/supp-info.
24. Science Journals: editorial policies. https://www.science.org/content/page/science-journals-editorial-policies.