CIRM
CALIFORNIA INSTITUTE FOR REGENERATIVE MEDICINE

Memorandum

To: Members of the Science and Finance Subcommittee
From: Janie Byrum, Ph.D., Senior Science Officer, Rosa Canet-Avilés, Ph.D., Chief Science Officer
Re: Data Science and Software Engineering (INFR9) Concept Plan
Date: May 29, 2026

Executive Summary

The proposed Data Science and Software Engineering Awards (INFR9) concept plan aims to help CIRM achieve the Strategic Allocation Framework goal of catalyzing the identification and validation of at least 4 novel targets and biomarkers by supporting the development of innovative, open source software (OSS) for integration of disparate datatypes to accelerate research in regenerative medicine. Such an award program will result in the development and deployment of broadly applicable, validated tools that solve multimodal data integration bottlenecks in stem cell-based and genetic therapy research. The CIRM team requests that the subcommittee recommends approval of the INFR9 concept plan to the full board, with a budget of $10M in research funds (award cap of $500,000 for a max duration of 2 years for each award) for FY 26-27.

I. Background

Biomedical research produces a massive volume and diversity of data, and there are ongoing struggles with data integration and interoperability (i.e., the ability to link data from different sources in a meaningful way). As a result, the full value of both individual datasets and modern analytical approaches remains underrealized. Across academia, industry, and research funding organizations, there is growing recognition that open source software tools—software that is transparently developed, freely available, reusable and extensible without restrictions—are essential infrastructure for overcoming these barriers. Tools built according to open source principles exhibit increased accessibility, reduced duplication of effort, and are more sustainable long-term. Scientific open source software enables researchers to connect heterogeneous datasets, apply robust analytical methods, and reproduce findings across systems and diseases. Recent community discussions, including national roadmaps for biomedical

data science and reports from interdisciplinary workshops, consistently emphasize the need for integrative tools capable of linking disparate data types.

At the ReMIND Program Meeting in October 2025, CIRM surveyed awardees about bottlenecks to adopting cutting-edge data science approaches in their labs. Respondents underscored the need for funding to support data scientists or computational personnel, user-friendly tools that enable data science for non-specialists, and resources to create or maintain open-source software for biomedical research. They also highlighted the importance of establishing data standards to improve interoperability, which reinforces the need for robust integrative software tools.

CIRM has made sustained investments in maximizing the value of CIRM-funded data and enabling data science. In 2023, CIRM implemented Data Sharing and Management Plans (DSMPs) for Discovery awards and has since implemented DSMPs across all R&D programs to track data outputs across the portfolio. To promote the findability of the datasets generated by CIRM researchers, CIRM launched the Data Explorer in 2025. Data Explorer currently includes over 800 public datasets and will continue to expand as additional datasets are deposited. Supporting tool development for data integration is the next step to expand the impact of these datasets and catalyze data science research.

By enabling connections across data types, platforms, and disease areas, software tools can accelerate the identification and validation of therapeutic targets and biomarkers. Open source tool development offers a democratized and leading-edge approach, ensuring that resulting tools are high quality, transparent, sustainable, and broadly reusable within the regenerative medicine community. Establishing an award program focused on open source software for data integration would allow CIRM to catalyze a next generation of computational resources that unlock the scientific impact of existing and future CIRM-funded data.

II. Proposal

Objective and Scope

The INFR9 program objective is to support the development of innovative, open source software for data integration of disparate datatypes to accelerate research in regenerative medicine. The expected outcome of an INFR9 award is the development and deployment of an open source software tool to solve a multimodal data integration bottleneck in stem cell-based and genetic therapy research that has wide applicability and scalable impact.

Investigator pairs will lead interdisciplinary projects focused on development, maintenance, and/or extension of software for data integration of disparate data types for target or biomarker identification and validation and/or for lead optimization to reduce time to therapeutic development and early clinical stages. To equip regenerative medicine labs with computational expertise, applicants must have both computational expertise and regenerative medicine expertise on the core team.

INFR9 awards will solicit a range of proposals to resolve bottlenecks in integrating data from different sources, including data from various techniques, model systems, and across molecular, cellular, tissue, organismal, and population scales. CIRM expects project teams to be resourceful in their open source software development and tool validation and leverage existing data, software infrastructure, automation, and community iteration to realize ambitious software projects that accelerate research. Each project will be required to justify how the proposed tool minimizes time and/or cost to identification and validation of therapeutic targets, biomarkers, or discovery of therapeutic candidates, as well as justify how the proposed tool provides an advantage over current tools.

Examples of INFR9 projects may include:

- Cell analytics software to readout present and future cell states and behavior without destructing the cells
- AI/ML approaches for in silico screening, protein design, molecule generation, cell and gene engineering
- Knowledge graphs for bridging structures, behavior assays, functional genomics, pathways, etc.

CIRM's Discovery programs allow tool development and data science activities when they directly support a research project that addresses a specific knowledge gap in stem cell or genetic research. However, these tools are typically ancillary components of individual research aims, with no requirements for software engineering best practices, open source development standards, or plans for community use or long-term sustainability. Maintenance and updates are dependent on an institution's priorities and staff turnover, and funding for computational staff within research labs is often fragmented. Tools developed as secondary components of a broader research effort with a lack of dedicated funding for software engineering together result in lower quality tools with narrow use cases and brief useful life.

INFR9 is distinct in its focus and structure. Rather than producing tools tightly tied to a single research project, INFR9 will support the development of broadly applicable, open source software designed for community use, scalability, and reproducibility. By requiring transparent, collaborative development and adherence to open source best

practices, INFR9 aims to generate tools that remain robust, well-maintained, and extensible beyond the originating team's priorities or personnel changes. This program directly addresses common pitfalls of lab-built software, such as a lack of documentation, limited maintainability, and dependence on individual trainees or collaborators, by establishing a structured pathway for sustainable software in regenerative medicine. The design of INFR9 demands a higher standard of tool utility and provides dedicated support for computational experts, strengthening the talent pipeline for research software engineering.

Novel Program Features

Considering INFR9 is product-focused (as opposed to research-focused) and its scope is to build data integration software solutions for research, this program incorporates several features that contrast with our existing research funding opportunities. The design and rationale for these features are also based on CIRM's experience managing discovery and infrastructure programs as well as feedback from open science pioneers and funders.

Deployment, Adoption, and Sustainability Plan -

To ensure that funded tools are broadly usable beyond research in the applicant's lab, the INFR9 program will require applicants to propose and justify a plan for deploying the software, facilitating adoption by a user base, and long-term maintenance. This requirement distinguishes INFR9 from software development that typically occurs within individual research labs, where tools are often created for a single project and lack sustained support. Applicants will be expected to outline plans for testing and quality assurance, deployment and release processes, issue tracking, user training and support, performance monitoring, feature updates, and communication of changes to the user community. Sustainability in INFR9 refers to the likelihood that tools are maintainable and have durable relevance to the field of regenerative medicine. Sustainability can be planned for by fostering an active user base that is motivated to contribute and maintain the code, by having clear documentation and guidelines for code maintainers, and by evangelizing the software project to researchers. Sustainability could also be planned for by raising money independently to fund open source work after the award via sponsors or crowdfunding, however fundraising is not an activity supported by INFR9.

Open Source Best Practices and Transparent Development -

To help advance open science in regenerative medicine and promote accessibility of data science tools, the INFR9 program will incorporate requirements for real-time public sharing of code, documentation, project plans, contributor guidelines, and more by requiring adherence to transparent development and open source software standards. This approach is the best practice for methods and tool development in life sciences, as it results in improved software products and greater knowledge transfer, and it accelerates the pace of

research. An example of a software project following these best practices is the Broad Institute's Genome Analysis Toolkit, which is a widely-used software toolkit for variant discovery in high-throughput sequencing data and is now maintained and extended by an active community user base. Its use cases have expanded from processing human exome data from one brand of sequencing instrument to genome data from any organism from any instrument.

Community and Contributor Engagement - Feedback and contributions from the user community are essential for developing the most appropriate tool and ensuring that it will be used by other researchers. INFR9 will require community engagement activities to solicit feedback on the software tool and inform its development to ensure relevance of the software tool to the field, improve its usability and accessibility, accelerate its development, align with emerging standards, and empower users to become contributors. Examples of these activities are hackathons, virtual demo sessions, and user feedback channels.

Data Sharing - In contrast to R&D awards (i.e., DISC, PDEV, CLIN2, RAPID), INFR9 awards will not support wet-lab research, and ergo will not be generating R&D data captured by CIRM Data Sharing and Management Plans (DSMPs). However, pre-existing datasets from wet-lab experiments will be used as a proof of concept or to validate the INFR9 software projects, and the INFR9 program will require that these datasets are freely available in the public domain at the time of application. This requirement is to ensure timely and predictable progress within the award period (i.e., prevent delays due to data generation), to provide the Grants Working Group (GWG) with a clear view of the proposed project, and to promote the reproducibility and re-usability of existing resources.

Award Mechanism and Parameters

The program will be implemented through a program announcement (PA) in FY 26-27 with a total budget allocation of $10M in research funds. Each award would be capped at $500,000 total costs with a max duration of 2 years. CIRM expects to fund 15-20 awards in the annual FY 26-27 application cycle.

III. Summary of Requested Action

The CIRM team requests that Science Subcommittee recommends approval of the Data Science and Software Engineering (INFR9) concept plan to the full board.

IV. Exhibit to Memo

- INFR9 Concept Plan