Documentation for Prepaire Labs Software Development

August 18, 2023

Sections:

1 Introduction of Prepaire Labs

Strictly Confidential 2023 Prepaire Labs Ltd

Prepaire Labs is a biotechnology company focused on drug discovery and regenerative medicine. Prepaire Labs' open architecture approach holds the potential to enable the development of new medicines and improving patient outcomes in more scalable ways.

It takes years for the pharmaceutical industry to create medicines capable of treating or curing human disease. Most current drug discovery is carried out by human chemists who rely on their knowledge and experience to select and synthesize the right molecules needed to become the safe and efficient medicines society depend on. To identify the synthesis paths, scientists often employ a technique called retrosynthesis – a method for creating potential drugs by working backward from the wanted molecules and searching for chemical reactions to make them.
Yet because sifting through millions of potential chemical reactions can be an extremely challenging and time-consuming endeavor, researchers at Prepaire Labs have created an AI framework called **HAiLO™** to automatically generate reactions for any given molecule. The new model shows that compared to current manual-planning methods, the framework is able to cover an enormous range of possible chemical reactions as well as accurately and quickly discern which reactions might work best to create a given drug molecule.

Using AI for things critical to saving human lives, such as medicine, is what we really want to focus on. Our aim was to use AI to accelerate the drug design process, and we found that it not only saves researchers time and money but provides drug candidates that may have much better properties than any molecules that exist in nature.

The **HAiLO™** model builds on a method that was able to generate molecule structures that exhibited desired properties better than any existing molecules.
Prepaire's team trained **HAiLO™** on a dataset that contains 100,000 chemical reactions provided by CAS and other sources. The framework "learns" from graph-based representations of given molecules, and uses deep neural networks to generate possible reactant structures that could be used to synthesize them. Its generative power can come up with hundreds of new reaction predictions in only a few minutes. By leveraging population-scale data, our platform constructs predictive models grounded in human genetic, phenotypic, and clinical data. These models provide insights into the underlying architecture and biology of diseases, facilitating the development of more accurate predictive models.

The generative **HAiLO™** model can supply multiple different synthesis routes and options, as well to rank different options for each molecule.
To further test the AI's effectiveness, Prepaire's team conducted a case study to see if HAiLO could accurately predict the best choice of molecules already in circulation or available “off-the-shelf”, against Covid-19 and Influenza COV2: Favipiravir, a broad spectrum oral anti-viral medication used to treat influenza; and Ivermectin, an anti-inflammatory medication which is used to treat river blindness. HAiLO was able to correctly generate a synthesis route for these medicines, and provided alternative combination synthesis routes that suggest more feasible and effective outcome. This led to the creation of a new combination drug called Cydalvym™ which was issued with two patents in 2022 and now going through pre-clinical trial Q3/4 2023.

Having such a dynamic and effective device at scientists' disposal could enable the industry to manufacture more effective drugs at a quicker pace – but despite the edge AI might provide scientists inside the lab, the medicines HAiLO or any generative AI creates still need to be validated.
Prepaire Labs utilizes induced pluripotent stem cells (iPSCs), genome editing and high-content cellular phenotyping to create in vitro disease models that optimize genetics, cell-type, environment, and multidimensional data collection for increased predictability of the human clinical outcomes.

Strictly Confidential 2023 Prepaire Labs Ltd

The power of Prepaire Labs was demonstrated with the release of **HAiLO**; this kind of product has been in the making by only a few small teams across the world.

Current drug-discovery platforms do not meet market constraints. The current competitors have embraced a closed technology approach, which will dramatically reduce their market reach. In that approach, the models are kept secret and is only served through client endpoint. This raises the following important concerns for businesses:

● Businesses wishing to use generative AI technology are forced to feed their valuable and sensitive to a black-box model, typically deployed in the cloud. This creates safety issues: models kept secret cannot be inspected to guarantee their outputs to be safe, thereby preventing them to be deployed in safety-critical applications. It also raises legal problems, in particular the one of falling under extraterritorial reach when sending sensitive data out of a company's or countries legal territory.

● Only exposing the output of models, instead of exposing the model entirely, makes it harder to connect with other components (retrieval databases structure inputs, and even clinical images). Hundreds of products are currently built by interconnecting model outputs and inputs to create composed capacities (memory, vision, etc.). Those products would work much better and faster were the models available
as white boxes.

● In some cases even the data used to train the model is kept secret, implying that we rely on a machine that has unidentified sources, and can produce uncontrollable outputs. Filtering efforts to address this issue are only a slim and breakable guarantee that the model will not output sensitive content on which it may have been trained.

Disrupting the market from the UAE

By creating the Prepaire operating system, we intend to train state-of-the-art models with counter-positions to closed-model current offerings. Our vision is to become a leading actor in the field, while developing a very valuable business around integrating these models in the pharmaceutical industry. Prepaire will become a provider of a ubiquitous medical operating system.
First demonstrating an end-to-end POC in the MENA market as a first step will create a defendable effort in itself — the UAE have already taken considerable steps to position themselves as a country willing to embrace NextGen technology.
Many talents in the various multi-disciplinary fields of expertise; as we have extensively tested, are willing to relocate to the UAE.

The Prepaire App Store is a cutting-edge platform specifically designed to catalyze advancements in bioinformatics and CRISPR technologies, particularly in the domains of drug discovery and regenerative therapies. This remarkable store presents a vast collection of applications and tools engineered to aid scientists and researchers in their pursuit of next-generation therapeutics.

Each tool and application within the Prepaire platform is fine-tuned to meet the complex demands of bioinformatics and CRISPR research. The tools cover a wide range of needs, from genome sequencing and analysis, drug-target interaction modelling, to CRISPR design and efficacy prediction. All tools are powered by the latest AI technology, automating and speeding up numerous intricate tasks that traditionally require substantial computational resources and time.

Prepaire App Store has the potential to completely transform the landscape of drug discovery and regenerative therapies. By simplifying and automating intricate processes, it can expedite the development of new drugs and regenerative treatments, bringing life-saving solutions to patients more rapidly. Moreover, its user-friendly design democratizes bioinformatics and genetic

engineering, allowing scientists at all levels, even those without extensive computational biology background, to contribute to and innovate within the field.
Additionally, the App Store promotes global collaboration by offering a platform for researchers to share findings, ideas, and unique approaches to problems. This collective intelligence can speed up the breakthroughs in managing complex health conditions and genetic disorders.

The Prepaire App Store has monumental implications for the future. The accelerated pace of drug discovery and development of regenerative therapies could lead to more effective responses to current and emerging health crises. By harnessing the power of CRISPR technology and making it accessible to a broader community, the platform could play a vital role in developing treatments for currently incurable diseases, ushering in a new era in medicine and global health.

Business development
On the business side, we will provide the most valuable technology brick to the emerging drug discovery-as-a-service industry that will revolutionise workflows.
We will co-build integrated solutions with best in class integrators based on industry requirements, and get extremely valuable feedback from this to become the main tool for drug discovery usage.
Integration with verticals can take different marketing forms, including licensing full access to the models, specialisation of models on demand, partnering with integrators/consulting companies to establish commercial contracts for fully integrated solutions.

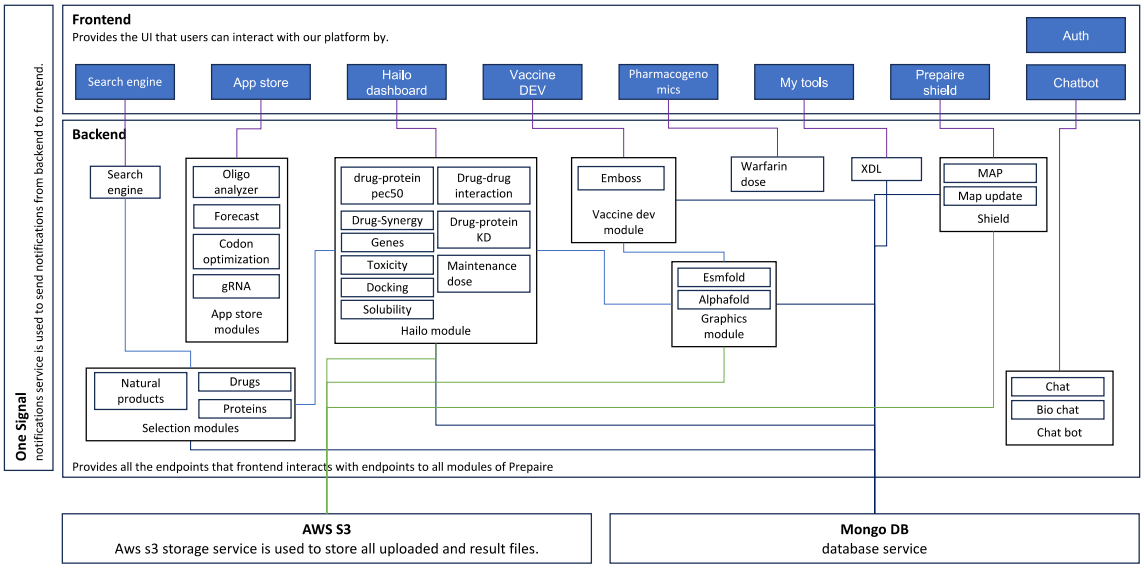
Infrastructure and data sources
Training a competitive model requires at least an exa-scale cluster for a few months. We intend to acquire this capacity. We have already negotiated competitive deals for acquiring our first Nvidia Superpod. Having trained models at large-scale before has provided us know-hows that will allow us to gain a factor 10-100 in training efficiency compared to public methods — our founders and early employees know exactly what to do to train the strongest model for a given computational budget.
Our early strategic partners are also biochemical data providers, and will open all necessary doors for acquiring high quality datasets on which our model can be trained and fine-tuned.

Roadmap
First year 2023
We will complete two pre-clinical trials before year-end, while developing business integration in parallel. The first generation of Prepaire will be partially open-source and rely on technology well mastered by the team. It will validate our platform efficacy and collaboration with 3rd party partners like Chemify who will be producing novel Oximes derived from the HAiLO™ discovery tool. The process will determine and measure our competence near partners, investigators and clinical institutions.
The second track is an off-the-shelf molecule combination therapy also generated from the HAiLO™ discovery tool. The outcome will address the significant shortcomings of current rapid drug discovery, repurposing and deployment methods, in response to sudden pathogenic outbreaks such at COV-2. This combination drug named Cydlavym was also awarded with two USPTO patents during 2022.

Next stages
One of the North stars of Prepaire Labs will be data security: we will release models in a well-staged way, making sure that our models can only be used for purposes aligned with our values. We will thus convince major public and private institutions to trust us for constructing the safe, controllable and efficient technology that we need to make humanity benefit from this science breakthrough.

2 Current Architecture

Prepaire platform architecture overview (As IS draft)

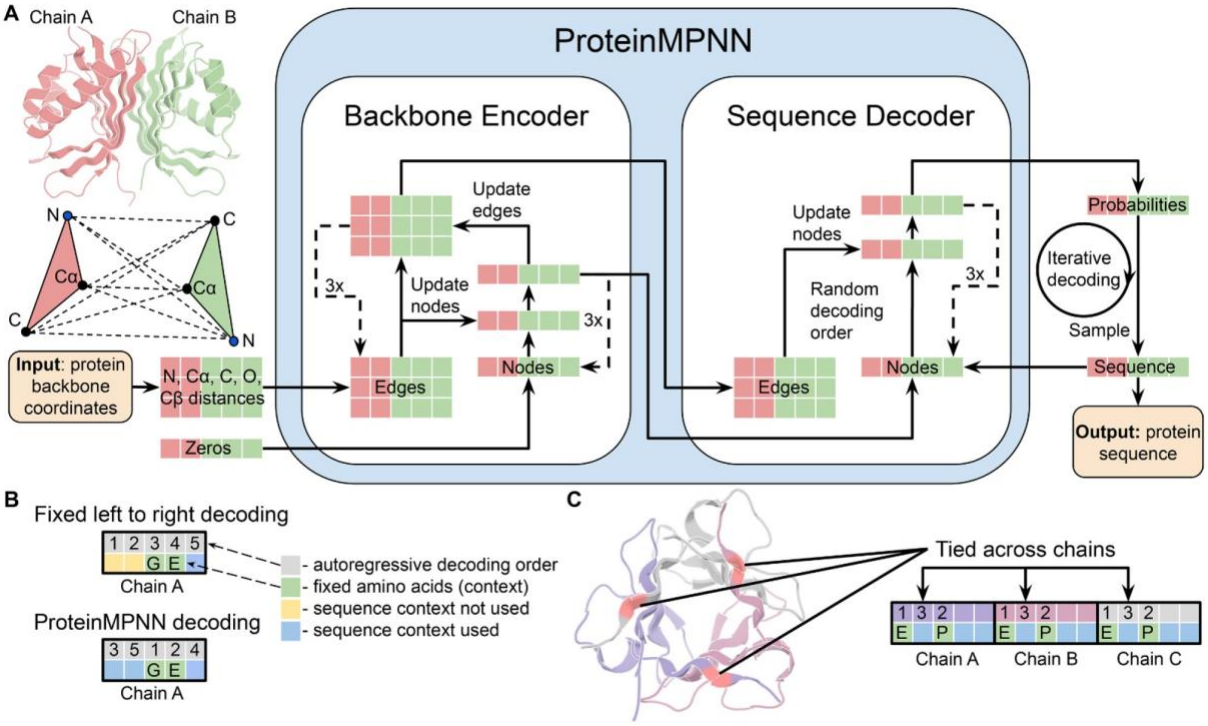


- **Data Collection and Preprocessing:**
  - Molecule Data: Collection of molecular structures, protein targets, and their properties.
  - Preprocessing: Convert molecular structures into a suitable representation (e.g., SMILES) for the neural network.

- **Generative Adversarial Networks (GANs):**
  - Generator: Utilizing deep learning to create novel molecular structures aiming to mimic real molecular data.
  - Discriminator: Differentiating between real and generated molecules, providing feedback to the generator.
  - Application: Identifying potential T cell receptors and generating new chemical entities for drug development.

- **Deep Neural Networks (DNNs):**
  - Architecture: Multiple hidden layers to model complex relationships between molecules and protein targets.
  - Training: Utilizing labeled data to train the network to predict the binding affinity between molecules and targets.
  - Application: In silico validation of targets, speeding up the drug discovery process.

Strictly Confidential 2023 Prepaire Labs Ltd



- **Neural Language Models (NLM):**
  - Modeling Chemical Language: Utilizing NLM to understand and generate chemical language (e.g., SMILES notation).
  - Integration with GANs: Collaborating with the generator to create more chemically valid structures.
  - Application: Enhancing the retrosynthesis process by providing human-like understanding of chemical structures and reactions.

- **Integration and Parallel Modeling:**
  - Parallel Processing: Running multiple models in parallel to explore various aspects of T cell receptor programming and molecule-protein interactions.
  - Integration: Combining insights from GANs, NLM, and DNN to provide a comprehensive view of potential drug candidates.

- **Validation and Testing:**
  - In Silico Testing: Utilizing computational methods to validate the predicted outcomes.
  - Wet Lab Integration: The UAE BSL3 and/or partnering labs

- **User Interface and Accessibility:**
  - Platform Integration: Integration into your existing "white box" platform to make this framework accessible to other scientists.
  - Interoperability: Ensuring compatibility with the Multi-Omics app store for seamless collaboration.

- **Compliance and Ethical Considerations:**
  - Data Security: Ensuring the privacy and security of the utilized data.

Strictly Confidential 2023 Prepaire Labs Ltd

- Ethical Guidelines: Adhering to international guidelines and regulations regarding gene editing and other biotechnological practices.

- **Ongoing Maintenance and Improvement:**
  - Updates: Regular updates to incorporate the latest advancements in AI and computational biology.
  - Scalability: Designing the system to easily scale with the growing needs of Prepaire Labs.

3 App Store Brief

Creating an operating system with an app store for bioinformatics, gene editing tools and protein prediction tools, integrated into a drug discovery platform Prepaire™

Conceptualization

- Define the Core Purpose: Understand the essential characteristics that your OS must have. This would include ease of use for bioinformatics, gene editing tools like Crispr Cas9, and protein prediction tools like Alphafold.
- Research the User Needs: Identify the needs of the users and developers in terms of the functionality, interface, and performance of the OS.
- Feature List: Create a list of features to be incorporated into the OS such as computing power, memory, security, data analysis tools, modeling tools, and the ability to interface with other systems.

Design

- UI/UX Design: Create an intuitive and user-friendly design. Since the intended audience is researchers and scientists, ensure that the design is simple yet professional, with an emphasis on functionality.

- System Architecture: Design the overall structure of the OS. This would include deciding on the kernel, device drivers, system utilities, and other aspects of the architecture.

- App Store Design: Create a blueprint for the app store, detailing how apps will be displayed, reviewed, downloaded, and updated.

Development

- Core OS Development: Start developing the kernel and other low-level features of the OS. This will be a complex task requiring expertise in system programming.

- App Store Development: In parallel, begin creating the app store. This will require a system for managing app submissions, updates, user reviews, and a secure payment processing system.

- Development of Support Tools: Create the necessary SDKs and APIs for developers to create and test their apps on your platform.

- Integration of LaaS: Connect your operating system with the Lab as a Service (LaaS) backend.

Testing and Validation

- Alpha Testing: Conduct internal testing to catch major bugs and issues.

- Beta Testing: Release the OS to a select group of users for beta testing. Collect feedback and make necessary adjustments.

- Compliance Validation: Verify if the OS meets all the necessary regulatory and legal requirements, especially given that it will handle potentially sensitive bioinformatics data.

4 Sample Journey

Whole Genome Sequencing (WGS) at 30x or 100x coverage offers a comprehensive genetic perspective on a patient. The following is a general protocol for analyzing patient data derived from WGS. This assumes that you are starting with raw data in the form of a FASTQ file and ends with a report containing the identified genetic variants and their possible clinical significance.

1. Quality Control:
The first step in any sequencing analysis should be a check of the quality of the raw data. Tools like FastQC can be used to analyze the quality of the FASTQ files.
Tools:

- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ FastQC
- https://github.com/s-andrews/FastQC

2. Alignment to Reference Genome:
Align the sequence reads to a reference genome using a tool like BWA or Bowtie. The output of this step is typically a BAM file.
Tools:

- https://bio-bwa.sourceforge.net BWA
- https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/04_alignment_using_bowtie2.html Bowtie

3. Post-Alignment Processing:
After alignment, the BAM file needs to be processed further. This usually includes sorting the file (Samtools), marking duplicate reads (Picard), and recalibrating the base quality scores (GATK BaseRecalibrator).
Tools:

- http://www.htslib.org Samtools
- https://github.com/samtools/samtools
- https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard- Picard
- https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrator GATK BaseRecalibrator

4. Variant Calling:
Call variants using a tool like GATK's HaplotypeCaller or Freebayes. This will result in a raw VCF (Variant Call Format) file.
Tools:

- https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller GATK HaplotypeCaller
- https://github.com/freebayes/freebayes Freebayes
- https://samtools.github.io/hts-specs/VCFv4.2.pdf VCF sample format

Strictly Confidential 2023 Prepaire Labs Ltd

5. Variant Quality Score Recalibration (VQSR):
This step uses machine learning to identify likely true variants and filter out false positives. This can be done using GATK's VariantRecalibrator and ApplyVQSR tools. After VQSR, you will have a recalibrated VCF file.
Tools:

- https://gatk.broadinstitute.org/hc/en-us/articles/360036510892-VariantRecalibrator GATK VariantRecalibrator
- https://github.com/broadinstitute/gatk/blob/master/src/main/java/org/broadinstitute/hellbender/tools/walkers/vqsr/ApplyVQSR.java
- https://gatk.broadinstitute.org/hc/en-us/articles/360037056912-ApplyVQSR ApplyVQSR

6. Annotation:
Annotate the VCF file using a tool like ANNOVAR or SnpEff. This adds information about each variant's known or predicted effects, which can help in interpreting the results.
Tools:

- https://github.com/WGLab/doc-ANNOVAR ANNOVAR
- http://pcingola.github.io/SnpEff/ SnpEff

7. Interpretation and Report Generation:
The final step is to interpret the annotated variants in the context of the patient's phenotype. Tools like InterVar can be used to apply the American College of Medical Genetics and Genomics (ACMG) guidelines for the interpretation of sequence variants. A detailed report should be generated, including all clinically significant findings.

Remember to comply with local and national regulations regarding genetic data handling and patient privacy. Some steps in this protocol may require significant computational resources, and bioinformatics expertise is essential for the correct interpretation of the results.
Tools:

- https://wintervar.wglab.org InterVar

Strictly Confidential 2023 Prepaire Labs Ltd

prepaire

prepaire SEARCH ENGINE APP STORE HAILO VACCINE DEV PHARMACOGENOMICS MY TOOLS PREPAIRE SHIELD

Back to App Store

DNA SEQUENCING & CRISPR

PATIENT

Name

Jane Doe

Date of birth

01/10/55

Sex

Female | Masculine

Race

Caucasian

INPUT DATA ACCEPTED

Sequence Type

DNA | Amino Acid

Product Type

Gene | GBlock | Megamar

Organism

Homo Sapiens (Human)

Sequence

GAGAGAATCGGAGAATCCCACAGCCAGATCAGGGACAGGAGAGATCC
CAGAAGGATCCTTGGAGGAAGACGAGGAGGAGGAGGAAGAGGATGTG
GGCTGGGAGGTCTATGGAGTTGGGGGCCCCATGGCCATGCTGACCCA
CCCTCATTCTTTTTCCAG

Choose a File

+ Add demographic

GENERATE NEW SEQUENCE | CLEAR | SUBMIT

Cell | Protein

PHENOTYPE VISUALISATION

Subject: Human (Female)

Age: 32 years

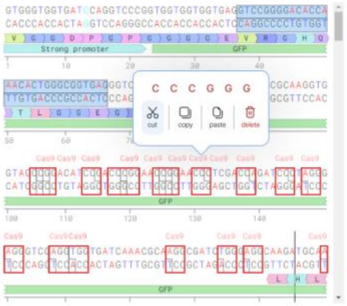
APPLY FOR VISUAL CONSENT | VISUALIZE OBJECT

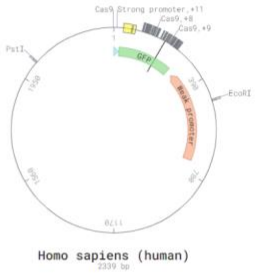
RESULTS SUMMARY

GTGGGTGGGGATCCTGGGCCTGGCGGGGGTGGGGAGGTGAGGGGACATCAGACCCTGGGAGGGGAGGGGAGGGTAGACACAGGGACCC
AGACCCAAGGTGGTTCACGGACCTCAGACCCAGAGAGGGGTGGCACAGGGACCCTAGACCAGATCCCCAGGGAGGGGAGGGGAGGAGAC
CAGACCCAGGGAGACCTGGGAGGCAAGATGCAGGGACCCCAGACCCAGGGTGGCCCAGGGATTCCAGACCCAGAAAGAGGAGGCAGAG
GGAGCCCAGATCTAGGAAGGCCCAGGGAGGAGAGGTGTGGCCCAGGAAGAGGCCCAGGTACCCCAGAACTGGGCTGAGAAGATATGGAGA
CCCTGCATTCCAATCTCTAAACTGAAACCTCTACCTCCAGACCCAGGAGGACCCAGGGATTCTAGACTCACAAAGGGGAGGCACAGGGACTC
CAAATCTAAGAAGGCCCAGGGAGGAGAAGCCCAGGGATTCCAGACCCAGAGAGGAGAGGCGTGGACACGCTAGACCCAGAAATTGGCTCA

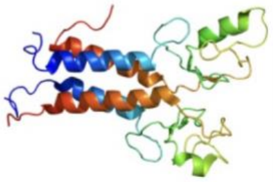
View all

A. MONOGENIC DISEASE RISK: 2 VARIANTS IDENTIFIED This test identified 2 genetic variant(s) that may be responsible for existing disease or the development of disease in this individuals lifetime. There are 2 primary genes linked with most families who have HBOC: BRCA1 and BRCA2. BRCA stands for BReast CAncer. A "mutation," or harmful genetic change, in either BRCA1 or BRCA2 gives a woman an increased lifetime risk of developing breast and ovarian cancers.

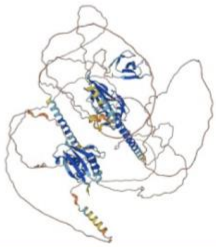




Old



New



Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation
Explanation

BioXP Video | PRINT DNA/RNA | ORDER IPSC | GENERATE REPORT

Strictly Confidential 2023 Prepaire Labs Ltd

Phase 2:
Once you have edited your target genes using a CRISPR system (like Cas9 or Prime Editing), the next steps involve packaging these edited genes for delivery, testing them ex vivo (outside the living organism) or in vivo (within the living organism), and verifying the changes.
Tools:

- https://github.com/topics/crispr-cas9?o=desc&s=updated Crispr
- https://github.com/uzh-dqbm-cmi/PRIDICT Prime editing

1. Packaging and Delivery:
You'll need to encapsulate the edited gene sequences for effective delivery to the target cells. This can be done using a variety of methods:

Viral vectors: Lentiviruses, adenoviruses, or adeno-associated viruses (AAVs) are commonly used to package and deliver edited genes.

Lipid nanoparticles (LNPs): These are lipid-based carriers that can encapsulate the RNA or DNA payload, protecting it from degradation and facilitating its entry into cells. To use LNPs, the edited genes need to be transcribed into RNA or packaged into plasmids.

Microinjection or electroporation: These physical methods can be used to directly introduce the genetic material into the cells.

2. Ex vivo or In vivo Testing:
Ex vivo: You can test the effectiveness of the gene edits in a controlled environment like a Petri dish. Introduce the packaged edited genes into cultured cells and then allow some time for the changes to take effect.

In vivo: For in vivo testing, the packaged edited genes are delivered into an organism. Depending on the study's goal, this could involve local (targeted to a specific tissue) or systemic (through the bloodstream) delivery.

3. Verification of Gene Editing:
Regardless of whether testing is done ex vivo or in vivo, the final step is verifying that the intended edits have occurred. This is typically done through:

PCR and Sequencing:PCR can be used to amplify the edited region, and then sequencing (Sanger or next-generation sequencing) can confirm the precise genetic changes.

Gene Expression Analysis: Techniques such as qPCR or RNA-seq can be used to verify changes in the expression of the edited gene.

Protein Level Analysis: Western blot or immunofluorescence can confirm changes at the protein level.

4. Functional Assays:
Finally, you would perform functional assays to determine the impact of the gene edits. The specific assays will depend on the gene of interest and the phenotype you expect it to affect.

5 SDK Brief

Overview
The goal of this project is to develop a Software Development Kit (SDK) to enable seamless integration of bioinformatics applications, gene editing tools, and protein prediction software into the "Prepaire" Drug Discovery Platform. Additionally, the SDK should also facilitate integration with our Prepaire App Store and support Codebot auto-coding functionality.

Objective
The primary objectives of this SDK are:

1. Create a standardized interface for third-party developers to integrate bioinformatics, gene editing, and protein prediction applications into the Prepaire platform.
2. Support integration with the Prepaire App Store to allow third-party applications to be distributed via our platform.
3. Enable Codebot auto-coding to streamline and automate the coding process during integration.

SDK Features
The SDK should include the following features:

1. Data Handling: The SDK should provide a robust set of tools for handling and processing genomic, proteomic, and other bioinformatic data types, including but not limited to sequences, alignments, and phylogenetic trees.

2. Algorithm Integration: The SDK should enable the use of common bioinformatics algorithms for tasks like sequence alignment, genome assembly, protein folding prediction, etc.

3. Application Integration: The SDK should provide APIs to facilitate easy integration of third-party applications with the Prepaire platform.

4. App Store Integration: The SDK should provide functionality to integrate the applications with the Prepaire App Store for distribution, including uploading, downloading, updating, and reviewing apps.

5. Codebot Auto-coding: The SDK should support the use of Codebot auto-coding, providing necessary hooks and endpoints for Codebot to work.

6. Testing and Debugging Tools: The SDK should provide a comprehensive set of tools for testing and debugging integrated applications, ensuring they work seamlessly within the Prepaire platform.

7. Documentation and Tutorials: Comprehensive documentation for all SDK features should be provided, along with tutorials to guide third-party developers in using the SDK.

Technology Stack
Given the nature of the SDK, it is recommended to use a widely adopted programming language with strong support for bioinformatics such as Python. Python has a broad range of libraries and frameworks for bioinformatics and machine learning which can be utilized in this SDK. Furthermore, Python is known for its simplicity and readability, which makes it easier for third-party developers to adopt.

Deliverables
1. The completed SDK with all the specified features.
2. Comprehensive documentation and tutorials to guide third-party developers in using the SDK.

Strictly Confidential 2023 Prepaire Labs Ltd

3. Example integrations with popular bioinformatics, gene editing, and protein prediction applications.
4. A robust set of testing tools and a debugging console to aid in the application development process.
5. Compatibility with Codebot auto-coding.
6. Integration of the SDK with the Prepaire App Store for easy app distribution.

Timeline
The estimated timeline for this project is 6-9 months, given the complexity of the software to be developed. This includes design, implementation, testing, and documentation phases.
The successful completion of this project will provide Prepaire a competitive advantage by enabling us to expand our platform's capabilities and to host a broader range of bioinformatics tools for drug discovery, thereby creating a richer ecosystem for our users and developers alike.

6 LIM Brief

- **User Interface (UI) Layer:**
  - Web-Based Interface: Allows scientists, researchers, and administrators to access the system from various devices.
  - Multi-Omics App Store: An integrated app store with gene editing applications.

- **Application Layer:**
  - Workflow Management: Manages the workflow of the samples, experiments, IPSC reprogramming, genome sequencing, etc.
  - Data Analytics and Visualization: Tools to analyze and visualize data from in silico and in vitro experiments.
  - Integration with Existing Tools: Facilitates the integration with other systems, gene editing applications, and third-party tools.

- **Data Management Layer:**
  - Database Management System (DBMS): Secure and robust database(s) to store experimental data, metadata, sample tracking, etc.
  - Data Interoperability: Ensures data compatibility with other systems and complies with standards like FAIR (Findable, Accessible, Interoperable, Reusable).

- **Security and Compliance Layer:**
  - Access Control: Role-based access control for secure user access.
  - Audit Trails: Maintains detailed logs of all system activities.
  - Compliance with Regulations: Ensures that the system adheres to legal and regulatory requirements, such as GDPR, HIPAA, etc.

- **Integration with Wet Lab (BLS 3 Lab):**
  - Sample Tracking: Real-time tracking of samples within the lab.
  - Equipment Integration: Connection with laboratory instruments for seamless data capture.
  - Inventory Management: Tracking and managing chemicals, reagents, and other lab resources.

- **Cloud & Infrastructure Layer:**
  - Scalable Architecture: Ensures that the system can scale as per the growing needs of Prepaire Labs.

Strictly Confidential 2023 Prepaire Labs Ltd

- Backup & Disaster Recovery: Regular backups and a solid disaster recovery plan.

- **API Layer :**
  - Open APIs: This could be beneficial if we want to allow third-party developers to create applications or integrations with our system.

The architecture should be flexible enough to accommodate our "white box" approach, allowing visibility and control over various processes, and fostering collaboration and innovation in the drug discovery and disease modeling sector.

7 Biocode Brief

Blockchain-based Biosecurity Control System for Biological and Genetic Sequence Printing

Abstract:
The present invention pertains to a novel biosecurity control system based on blockchain technology designed to regulate and monitor the printing of genetic sequences. The system integrates the principle of distributed ledger technology to establish a secure and transparent method for managing permissions for genetic sequence printing tasks, thus significantly enhancing biosecurity measures.

The blockchain system stores immutable records of all authorized printing tasks, including the details of the genetic sequences, the verified parties involved, the timestamps, and the granted permissions. The secure nature of the blockchain ensures that the stored information cannot be altered or deleted, and is readily available for audit and review.

A unique feature of this system is its ability to prevent unauthorized printing tasks. Without a corresponding record of permission on the blockchain, verified and approved by an authorized party, the printer will not initiate the genetic sequence printing process. This serves as an effective deterrent against unregulated and potentially hazardous genetic sequence printing.
In essence, this blockchain-based biosecurity control system establishes an innovative, reliable, and secure approach for managing and regulating genetic sequence printing, thereby fostering a safer bioengineering environment.

Introduction:
A blockchain ledger is a decentralized and immutable record-keeping system that can revolutionize the management of biological printers. In the context of biological printers, blockchain technology can enhance security, transparency, and traceability in the printing process.

- Blockchain Technology:
  - Blockchain is a distributed ledger technology that securely records and verifies transactions across multiple network participants.
  - It consists of blocks of data linked together in a chronological chain, creating an immutable and tamper-resistant record.

- Benefits of Blockchain for Biological Printers:
  - Security:
  - Blockchain utilizes advanced cryptographic techniques to ensure the security and integrity of data.
  - It reduces the risk of unauthorized access, data manipulation, and counterfeiting, enhancing trust in the biological printing process.

- Transparency:

Strictly Confidential 2023 Prepaire Labs Ltd

- Blockchain provides transparent and auditable records of every transaction within the biological printing network.
- All participants have access to a shared, real-time view of the data, promoting accountability and eliminating information silos.

- Traceability:
  - Blockchain enables end-to-end traceability of biological printing activities.
  - Each print job, including materials used, printing parameters, and quality control measures, is recorded on the blockchain, allowing easy verification and tracking.

- Workflow of Blockchain-based Biological Printing:
  - Job Creation:
  - Users initiate print jobs by submitting design files and specifying requirements.
  - Details such as desired materials, quantity, and printing parameters are recorded and stored on the blockchain.

- Material Sourcing and Verification:
  - Blockchain ensures that only verified and approved materials are used for biological printing.
  - Suppliers and manufacturers can record their materials' origin, quality, and certifications on the blockchain, providing transparency and trust.

- Printing Process:
  - As the printing process begins, each step, including calibration, layering, and post-processing, is logged on the blockchain.
  - Timestamped records ensure an accurate and auditable history of the printing process.

- Quality Control:
  - Quality control measures, such as inspections, tests, and certifications, are logged on the blockchain.
  - Users can easily verify the quality of printed products and identify any potential issues or recalls.

- Smart Contracts and Automation:
  - Smart contracts, programmable self-executing agreements, can automate various aspects of the biological printing process.
  - They enable automatic payment processing, material ordering, and compliance checks, streamlining operations and reducing human intervention.

Conclusion:
Implementing a blockchain ledger for biological printers offers numerous advantages, including enhanced security, transparency, and traceability. By leveraging blockchain technology, the biological printing industry can promote trust, efficiency, and quality control in this innovative field.

Development:

Using ChatGPT for Smart Contract and Blockchain:
ChatGPT, a large language model developed by OpenAI, has the potential to revolutionize smart contracts and blockchain technology. ChatGPT's AI capabilities can enhance the accuracy and efficiency of smart contract execution, improve smart contract coding, enhance blockchain security, and aid in analyzing and interpreting large amounts of data for blockchain applications

Strictly Confidential 2023 Prepaire Labs Ltd

Introduction:
The advancement of Artificial Intelligence (AI) has revolutionized various industries in recent years. ChatGPT, a large language model developed by OpenAI, is one of the latest AI innovations that can potentially transform how we think about smart contracts and blockchain technology.
Smart contracts are self-executing contracts that use blockchain technology to automate the execution of contract terms. In contrast, blockchain technology is a decentralized digital ledger that records transactions securely and transparently.
ChatGPT's AI capabilities can enhance the accuracy and efficiency of smart contract execution, improve smart contract coding, enhance blockchain security, and aid in analyzing and interpreting large amounts of data for blockchain applications.

Prepaire Labs will explore how ChatGPT's AI is changing smart contracts and blockchain technology and ChatGPT's role in automating and optimizing smart contract execution.
Smart contracts are designed to automate the process of contract execution, ensuring that contractual terms are met without the need for intermediaries. ChatGPT's AI capabilities can automate and optimize smart contract execution in several ways.

- Firstly, it can be used to streamline the process of contract creation by assisting in the development and testing of smart contract codes. Also, ChatGPT can help developers write more efficient and error-free code, reducing the likelihood of bugs and other issues.
- Secondly, ChatGPT's AI can be used to automate the process of contract execution. By analyzing and interpreting the data generated by smart contracts, ChatGPT can help identify potential issues or errors in the contract code, alerting developers to take corrective action. This can reduce the time and effort required to manually monitor smart contracts, improving the speed and accuracy of contract execution.
- Thirdly, ChatGPT's AI capabilities can optimize smart contract performance by analyzing contract data and identifying patterns and trends. This can help improve the efficiency of contract execution by identifying areas where the contract can be optimized, such as reducing gas fees or improving the execution speed.
- Improving smart contract coding with ChatGPT's language capabilities Smart contracts are coded using programming languages such as Solidity, which is specifically designed for writing smart contracts on blockchain platforms such as Ethereum. Other programming languages, such as Python, JavaScript, and C++, can also be used for smart contract coding. However, coding smart contracts can be complex and error-prone, as even small mistakes in the code can have significant consequences. This is where ChatGPT's natural language processing capabilities can come in handy. With ChatGPT's AI, developers can write smart contract code in natural language, which can help reduce errors and improve the efficiency of the coding process.

Enhancing blockchain security with ChatGPT's advanced threat detection and prevention
Blockchain technology has emerged as a highly secure and transparent way to store and transfer digital assets. However, blockchain is not immune to security threats as with any technology.
Some common security threats that blockchain technology faces include:
Hacking and Cyberattacks, Insider threats, and 51% Attacks.
ChatGPT's AI can help prevent and mitigate these threats in a number of ways.
ChatGPT can analyze network traffic and detect unusual activity, such as suspicious transactions or attempted hacks. It can also monitor social media and other sources to identify potential threats, such as discussions of vulnerabilities or attacks.
Additionally, ChatGPT can use machine learning to identify patterns of behavior that may indicate an insider threat and alert administrators to take action.
Analyzing and interpreting large amounts of data for blockchain applications

Strictly Confidential 2023 Prepaire Labs Ltd

Blockchain technology can be used for biological supply chain tracking and other data-intensive applications. By creating an immutable, decentralized ledger of transactions, blockchain allows for secure and transparent tracking of base compounds.
For example, in the supply chain industry, blockchain technology can track the movement of goods from the point of origin to the final destination. This generates a huge amount of data, including information about product quality, shipping times, and inventory levels. ChatGPT's AI can analyze this data, identify patterns and trends, and provide real-time insights that can help optimize supply chain operations. In conclusion, ChatGPT's AI capabilities have the potential to revolutionize the way we think about smart contracts and blockchain technology. By streamlining and automating smart contract execution, enhancing smart contract coding accuracy and efficiency, improving blockchain security, and enhancing blockchain data analysis,

Project brief

Objective: Creation of Blockchain system with ledger containing proof-records of all printed tasks and granted permissions for printing by authorized parties with the purpose of bio security control, so without explicit blockchain-recorded permission by verified party, the printer could not start printing the sequence.

Phase 1:
Business Analysis for Comprehensive Solution: The insights garnered from this analysis and feasibility study will inform the design of the comprehensive solution for Prepaire, ensuring it is appropriately tailored to meet Prepaire specific needs.

Key stakeholders needed for the successful completion of the project:
Prepaire Labs / Carl Freer - Product Owner - to provide neccesary access to key people and answer general questions on the whole ecosystem

Telesis Hardware Engineer and Architect of the BioXp system:
BioXp system currently communicates with the Telesis Bio server via LAN, making an outbound SSL connection (HTTPS/SSL) to logmein.com, telesisbio.com and drive.google.com.
The integration of blockchain with the physical bioprinter is a key part of this project. Most likely each printer will have its own encrypted private key that will sign transactions on blockchain to verify the device. So the architect of the current system will be required to understand the existing design and how blockchain private key management piece can be added to the existing printers to ensure that the bioprinters are capable of interacting securely with the blockchain network.
Hardware engineers who have developed the solution would be also nice to have to understand if there's a risk of the possible man-in-the middle hardware attacks vectors and points where the printing job can be altered by attacker who has direct access to the devices.
It will help the cybersecurity team to validate safety of the final solution.

Biotechnology/Bioengineering Experts:
These professionals can provide valuable insights into the functionality and security needs of the bioprinters, and they can also help validate the project's impact on the bioprinting process.
To design the final blockchain systems, these experts can provide valuable requirements on the average length of printed sequences, frequency of received tasks and duration of prints, making possible for the blockchain arhitect to estimate the volume of stored data and smart contracts structure.

BioXp System User Representatives:
These could be researchers, laboratory professionals, or anyone who would use the bioprinters in a practical setting. They can provide critical feedback on the system's usability

Strictly Confidential 2023 Prepaire Labs Ltd

and functionality and reveal relevant requirements on additional use cases that may be covered (for example, if job cancellations needs to be recorded on blockchain, any other non standard situations and use-cases in the printer life-cycle) .

Regulatory Experts:
International Gene Synthesis Consortium (IGSC) gene synthesis members apply to prevent the misuse of synthetic genes. By uniformly screening the sequences of ordered genes and vetting gene synthesis customers, IGSC members collaborate to establish and continuously improve best practices, safeguard the many benefits of gene synthesis technology while minimizing risk, and help ensure broad compliance with HHS Guidance for Double-Stranded DNA Providers and other international standards. Given the intersection of biotechnology and data security, there are likely to be complex regulatory issues to navigate some of which are mentioned in [https://files.telesisbio.com/docs/IGSC_Harmonized_Screening_Protocol.pdf].
Experts in relevant legal and regulatory areas will be necessary to ensure the final blockchain-enabled solution is compliant with all laws and guidelines.
It will help to collect requirements on how current systems for sequences validation are working (most likely they will need to be integrated into blockchain solution, or will be part of requirements to the blockchain solution API/SDK for early detection of misuse of the whole system).

Gene Sequence Screening System expert:
Will help to gather integration requirements on how Gene Sequence Screening Systems are working at the moment, what are the business processes, how many variations of these systems are out there, and how they can be integrated with the blockchain (if such integration is needed).

Customer Screening System expert:
Will help to gather integration requirements on how Customer Screening Systems are working at the moment, which user parameters are captured, how KYC is performed, how its stored and validated and who has access to it. This will help the blockchain team to come up with the architecture and integration requirements that will accomodate all pieces of needed information as well as processes and user roles/access rights to it.

Record keeping system expert:
IGSC members retain records of every gene synthesized and delivered for a minimum of 8 years after shipping, including at least the following: (a) the synthetic DNA sequence; (b) the vector (if applicable); and (c) the recipient's identity and shipping address.
For storing this information on the blockchain, requirements have to be gathered on existing business processes and existing systems. So expert is needed to help with this.

Reporting process Experts:
GSC members have established relationships with local and national law enforcement and intelligence authorities with whom we can share information to reportand to prevent the potential misuse of synthetic genes. This misuse needs to be captured on the blockchain to keep track-record of obusing parties. Expert who can help gather requirements on reporting process and systems are needed to automate process of recording such events in the blockchain.

Data Privacy Experts:
Given the sensitive nature of bioprinting and the potential privacy implications, these stakeholders can ensure the project complies with relevant data protection laws and guidelines. For example - requirements on secure encryption of stored printed sequences and definition on what types of users can access/decrypt/use/view these sequences.

Strictly Confidential 2023 Prepaire Labs Ltd

Bioethics Professionals:
Due to the nature of the data being handled (DNA sequences), experts in bioethics are crucial to ensure the project adheres to ethical standards and best practices. Potentially can reveal requirements on public access to some pieces of data stored on the blockchain so community can also validate some essential parts of the process/records.
Singularity Universe - Service Provider

Cryptographers/Cybersecurity Experts:
Ensuring the security of the blockchain system is crucial to protect against unauthorized access or hacking attempts. These experts can help design robust security protocols for the blockchain system and bioprinters, as well as foresee possible attacks vectors (including on private-keys infrastructure, hardware attacks that can substitute approved printing job into something different) and improve the final architecture security.

Blockchain Architect:
The team is essential to design, test, build, and maintain the distributed blockchain ledger. Their tasks would include architecting blockchain bioprinters ecosystem and developing smart contracts to manage permissions and creating interfaces between the blockchain and BioXp system (including printers and the Telesis Bio web server).

Project Manager:
Effective project manager will ensure that the project stays on track, within budget, and meets the defined objectives.

Business Analytic:
The BA will work closely with stakeholders to understand their needs and translate them into requirements.

Deliverables:
Feasibility Study Report: This comprehensive document outlines the feasibility of the proposed solution in terms of technical feasibility, economic feasibility, legal feasibility, operational feasibility, and scheduling feasibilit:

- Business Case: This is an argument, usually in a document, that helps the business decide whether the project is worth the investment. It includes a cost-benefit analysis, risks, assumptions, and strategic alignment with business goals.
- Requirements Documentation: In the discovery phase, initial high-level business, user, and functional requirements will be gathered.
- Stakeholder Analysis: The BA identifies who the stakeholders are, what their interests and concerns might be, and how they will interact with the proposed solution.
- Gap Analysis: This identifies the difference between the current state of the business and the proposed future state, outlining what needs to be done to move from the current to the desired state.
- Risk Analysis: This identifies potential risks in proceeding with the project, along with possible mitigation strategies.
- Project Plan: This high-level document outlines the project's goals, scope, deliverables, required resources, budget, and timeline.

Strictly Confidential 2023 Prepaire Labs Ltd

8 Instructions to Locate the Best AI Models for Drug Discovery, Cytotoxicity, and Pharmacogenomics

1. Define Objectives and Criteria:
   - Drug Discovery: Identify AI models that can predict novel drug candidates, molecular interactions, and drug efficiency.
   - Cytotoxicity: Find models that predict toxicity levels in different cells, enabling safe drug development.
   - Pharmacogenomics: Look for models that link genetic information with drug responses, aiding in personalized medicine.

2. Research Existing AI Models and Platforms:
   - Browse scientific literature, repositories, and platforms like GitHub, arXiv, and PubMed.
   - Focus on models with high accuracy, precision, and compatibility with Prepaire's LIM operating system.

3. Evaluate Model Performance:
   - Consider aspects such as prediction accuracy, scalability, and interpretability.
   - Determine if the models align with a "white box" approach, ensuring transparency and understanding.

4. Compatibility Check:
   - Ensure that the models can be integrated into Prepaire's Multi-Omins app store.
   - Verify support for gene editing applications and compatibility with the BLS 3 lab environment.

5. Connect with Model Developers and Communities:
   - Engage with AI researchers, developers, and communities in the medical and pharmaceutical sectors.

Strictly Confidential 2023 Prepaire Labs Ltd

- Utilize in silico connections with in vitro to align with Prepaire's objectives.

6. Perform Proof-of-Concept Testing:
   - Select promising models and run them on sample datasets.
   - Evaluate how well they fit into Prepaire's ecosystem and their impact on drug discovery, cytotoxicity, and pharmacogenomics.

7. Legal and Ethical Considerations:
   - Review intellectual property rights, privacy concerns, and other legal matters.
   - Consider ethical implications, particularly in gene editing and personalized medicine.

8. Documentation and Reporting:
   - Document findings, evaluations, and proof-of-concept results.
   - Create a comprehensive report to share with Prepaire's leadership, outlining recommendations and insights.

9. Implementation and Integration (if applicable):
   - Plan for integrating the selected models into Prepaire's platform.
   - Work closely with the development team to ensure seamless implementation.

Model Selection:
Deep Neural Networks (DNNs)

1. Residual Networks (ResNets): Great for image analysis, these networks have skip connections that help in training very deep networks without the vanishing gradient problem.

2. Long Short-Term Memory Networks (LSTMs): Particularly useful for sequential data like time-series, LSTMs can be applied in patient record analysis and predicting disease progression.

3. Transformer Models: Originally designed for NLP tasks, transformer models have been adapted for various purposes, including genomic sequence analysis.

Generative Adversarial Networks (GANs)

1. CycleGANs: Used for image-to-image translation tasks, CycleGANs can be utilized for medical imaging translation between modalities.

2. Wasserstein GANs (WGANs): Known for stability in training, WGANs could be used for generating synthetic medical images or genomic sequences.

3. BigGAN: Specializing in generating high-resolution and high-quality images, BigGAN could be useful for enhancing low-quality medical images.

Large Language Models

1. BERT (Bidirectional Encoder Representations from Transformers): BERT can be employed for extracting insights from clinical texts, electronic health records, and scientific literature.

2. GPT (Generative Pre-trained Transformer): GPT models, such as GPT-3 or 4, are highly flexible and can be fine-tuned for a variety of tasks including drug discovery, protein folding, and patient engagement through natural language interfaces.

Strictly Confidential 2023 Prepaire Labs Ltd

3. BioBERT: A version of BERT specifically pre-trained on biomedical texts, BioBERT could be powerful in extracting insights from scientific literature related to drug discovery and pharmacogenomics.

Considerations for Prepaire

1. Alignment with Objectives: Ensure that the selected models align with Prepaire's specific goals in drug discovery, cytotoxicity, and pharmacogenomics.

2. Customization and Integration: Assess how easily these models can be customized and integrated into the existing Prepaire platform and workflows.

3. Compliance and Ethics: Keep in mind the legal and ethical considerations, particularly concerning patient data privacy and intellectual property.

4. Benchmarking Strategy: Develop a robust benchmarking strategy to systematically compare these models against Prepaire's in-house solutions.

5. Collaborations and Partnerships: Consider collaborating with academic institutions, industry leaders, and AI research groups that have expertise in these models.

Locations & Links
Deep Neural Networks (DNNs)

1. Residual Networks (ResNets):
   - GitHub Repository: [Keras ResNet](https://github.com/keras-team/keras-applications/blob/master/keras_applications/resnet.py)
   - Available in popular deep learning frameworks like TensorFlow and PyTorch.

2. Long Short-Term Memory Networks (LSTMs):
   - TensorFlow Tutorial: [LSTM with TensorFlow](https://www.tensorflow.org/tutorials/structured_data/time_series)
   - PyTorch LSTM: Available through the standard PyTorch library.

3. Transformer Models:
   - Hugging Face Transformers: [Transformers Library](https://github.com/huggingface/transformers)
   - Also available in TensorFlow and PyTorch.

Generative Adversarial Networks (GANs)

1. CycleGANs:
   - GitHub Repository: [CycleGAN](https://github.com/junyanz/CycleGAN)

2. Wasserstein GANs (WGANs):
   - GitHub Repository: [WGAN](https://github.com/martinarjovsky/WassersteinGAN)

3. BigGAN:
   - GitHub Repository: [BigGAN](https://github.com/ajbrock/BigGAN-PyTorch)

Large Language Models

1. BERT (Bidirectional Encoder Representations from Transformers):

- Hugging Face Transformers: [BERT](https://huggingface.co/transformers/model_doc/bert.html)

2. GPT (Generative Pre-trained Transformer):
- Hugging Face Transformers: [GPT-2](https://huggingface.co/transformers/model_doc/gpt2.html)
[GPT-3](https://platform.openai.com/docs/api-reference/models/gpt-3.5-turbo)
GPT -4

3. BioBERT:
- GitHub Repository: [BioBERT](https://github.com/dmis-lab/biobert)

Onboarding Algorithms and Models to Prepaire Platform

1. Evaluation and Selection:
- Confirm the models meet Prepaire's requirements for drug discovery, cytotoxicity, and pharmacogenomics.
- Verify the compatibility with the LIM operating system and the Multi-Omins app store approach.

2. Obtain Necessary Permissions and Licensing:
- Secure the intellectual property rights or licensing agreements for using the selected models.
- Ensure adherence to legal, ethical, and regulatory standards.

3. Data Preparation:
- Organize and prepare the data that the models will be trained on or applied to.
- Ensure data compatibility and preprocessing as required by specific models.

4. Model Customization and Optimization:
- Adjust model parameters and configurations to align with Prepaire's unique platform.
- Optimize for performance, scalability, and integration with existing applications and services.

5. Development Environment Setup:
- Create development and testing environments that mirror the production setting.
- Install necessary dependencies, libraries, and tools required by the models.

6. Integration with Prepaire's Multi-Omins App Store:
- Develop connectors or APIs to integrate models into the app store.
- Ensure the models are accessible and functional within Prepaire's ecosystem.

7. Quality Assurance and Testing:
- Conduct rigorous testing to ensure the models are working as expected.
- Validate against various use cases and scenarios specific to Prepaire's goals.

8. Deployment and Scaling:
- Plan and execute a phased deployment to the production environment.
- Monitor performance and scalability to meet expected demands.

9. User Training and Support:
- Develop training materials, tutorials, and documentation for scientists and other end-users.
- Provide ongoing support to address questions, troubleshoot issues, and implement updates.

10. Monitoring and Maintenance:
- Implement monitoring tools to track performance, usage, and other key metrics.

Strictly Confidential 2023 Prepaire Labs Ltd

- Regularly review and update models to ensure they remain effective and aligned with Prepaire's objectives.

11. Collaboration with Wet-Lab Services:
- If applicable, coordinate with downstream wet-lab services like IPSC reprogramming and genome sequencing.
- Ensure seamless communication between in silico and in vitro components.

9 Framework for Collecting and Uploading Medical Data to Prepaire

A. Data Collection

1. Imaging Data:
- Utilize standard protocols for collecting medical imaging data (e.g., MRI, CT scans).
- Ensure proper formatting and consistency across different imaging sources.

2. Patient Data:
- Gather relevant patient information, such as demographics, medical history, and clinical notes.
- Anonymize data to remove or replace any identifying information (see Section B).

3. Genomic Data Sets (FAST Files):
- Collaborate with genomics labs and partners to collect genomic sequence data.
- Follow standard file formats (e.g., FASTQ) to ensure compatibility.

B. Data Anonymization

1. Implement De-identification Techniques:
- Remove all identifiable information like names, addresses, IDs.
- Use pseudonymization where needed, replacing identifiers with non-identifiable codes.

2. Apply Data Masking and Encryption:
- Apply masking to sensitive data fields.
- Encrypt data at rest and in transit to ensure security.

C. Data Preprocessing

1. Validate Data Quality:
- Check for consistency, completeness, and accuracy.
- Handle missing values and outliers as needed.

2. Standardize Formats and Structures:
- Convert data into consistent formats compatible with Prepaire's platform.
- This may include converting imaging files to specific formats or preprocessing genomic FAST files.

D. Data Upload to Prepaire

1. Secure Data Transmission:
- Use secure protocols (e.g., SFTP, HTTPS) for transmitting data.
- Ensure adherence to data protection regulations and compliance standards.

2. Integration with Prepaire's Platform:
   - Utilize APIs or direct integration to upload data into Prepaire's system.
   - Ensure compatibility with existing data storage and processing infrastructure.

3. Metadata and Data Cataloging:
   - Include relevant metadata for each dataset to facilitate data discovery and management.
   - Maintain a catalog of uploaded datasets with appropriate versioning.

E. Ongoing Data Management and Governance

1. Implement Access Control and Auditing:
   - Define user roles and permissions to control access to data.
   - Regularly audit data access and utilization to detect and prevent unauthorized activities.

2. Monitor Data Quality and Integrity:
   - Establish ongoing monitoring and quality checks.
   - Implement mechanisms for data updates, corrections, and maintenance.

3. Compliance with Legal and Ethical Guidelines:
   - Regularly review adherence to legal and ethical guidelines, including GDPR and HIPAA.

10 CD Library Synthesis Model

Below is the framework for a machine learning model that can be used to identify suitable clustered differentiations (CDs) for T cells and NK cells, to assist in the targeting of cancer cells. It a complex but rewarding task, considering the potential implications for personalized medicine:

1. Data Collection and Preprocessing:
   - CD Information: Collect information on all known CDs (approximately 400 or more). This information could include the structural, functional, and interaction characteristics.
   - Cell Data: Collect data on how various CDs interact with T cells and NK cells. This could include success rates in transfections, binding affinities, activation levels, etc.
   - Preprocessing: Clean and preprocess the data, normalizing and transforming where necessary to make it suitable for machine learning models.

2. Feature Engineering:
   - Feature Selection: Identify the critical features that will allow the model to discern patterns between different CDs and their suitability for targeting cancer cells.
   - Feature Creation: Create new features if necessary, using domain knowledge about molecular biology and immunology.

3. Model Selection and Training:
   - Model Architecture: Based on the nature of the data and problem, you might choose supervised learning algorithms like Random Forest, SVM, or deep learning models like Convolutional Neural Networks (CNNs), especially if structural data are involved.
   - Hyperparameter Tuning: Optimize the model parameters using techniques like grid search or random search.
   - Training: Train the model on a labeled dataset (if you have it) or create one based on expert knowledge.

4. Validation and Evaluation:

- Validation: Split the dataset into training, validation, and test sets to ensure robust evaluation.
- Metrics: Utilize metrics like accuracy, precision, recall, F1-score, or custom metrics tailored to your specific needs.

5. Interpretation and Visualization:
- Feature Importance Analysis: Utilize tools like SHAP to understand the impact of different features (such as specific structural or functional characteristics of CDs).
- Visualization Tools: Provide visualization to interpret the results, allowing researchers to understand the underlying patterns better.

6. Integration with Prepaire Platform:
- API Development: Create an API for the model to be integrated into your platform, allowing users to input specific parameters and receive recommendations for suitable CDs.
- User Interface: Design an intuitive UI that allows users to interact with the model, offering insights and suggestions for the best "armory."

7. Continuous Improvement:
- Feedback Loop: Implement a system that collects feedback from users and real-world results, refining the model over time.
- Monitoring: Regularly monitor the model's performance, update with new data, and retrain if necessary.

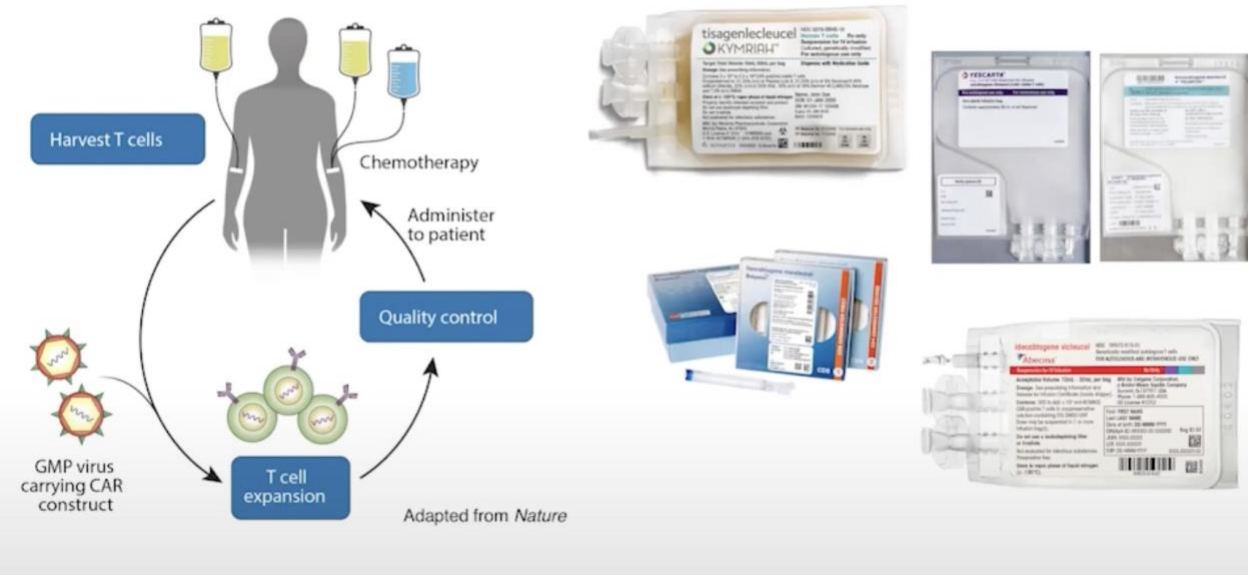
11 CAR T Receptor Model

Alex Marson is the director of the Gladstone UCSF Institute of Genomic Immunology and a professor at UCSF. He started his lab as a Sandler faculty fellow at UCSF in 2013.
In addition to these positions, he's a scientific director of human health at the Innovative Genomics Institute, a member of the Parker Institute for cancer immunotherapy, and an investigator at the Chan Zuckerberg Biohub. His research is focused on adapting CRISPR genome editing techniques to human immune cells to understand the genetic programs controlling immune cell function and to manipulate T-cells, to generate cell-based therapies.

I want Prepaire™ platform users take advantage of gene engineering technologies, to not only understand specific DNA sequences, that control critical functions of cells in the human immune system, but also to think about how we can take those lessons and start reprogramming human immune cells to have new functions that transform them from cells into cell therapies for a range of human diseases. We are broadly inspired by successes in the world of cell therapy, which in some cases have become FDA approved. These cell therapies that are being clinically delivered are genetically modified human immune cells. The therapies which are farthest along are from the field of CAR T-Cells, and this is for certain types of cancer.

The current CAR T-Cell patient's immune cells are taken out of circulation outside of the body, and then genetically modified. They're currently modified with viral vectors that insert genetic material into non-targeted sites in the genome of those T-cells. The newly programmed transfected artificial receptor, (a Chimeric antigen receptor) directs the T-cells against targets on certain types of cancer. And then these genetically modified T-cells could be reinfused into a patient. The results for certain cancers have been remarkable.

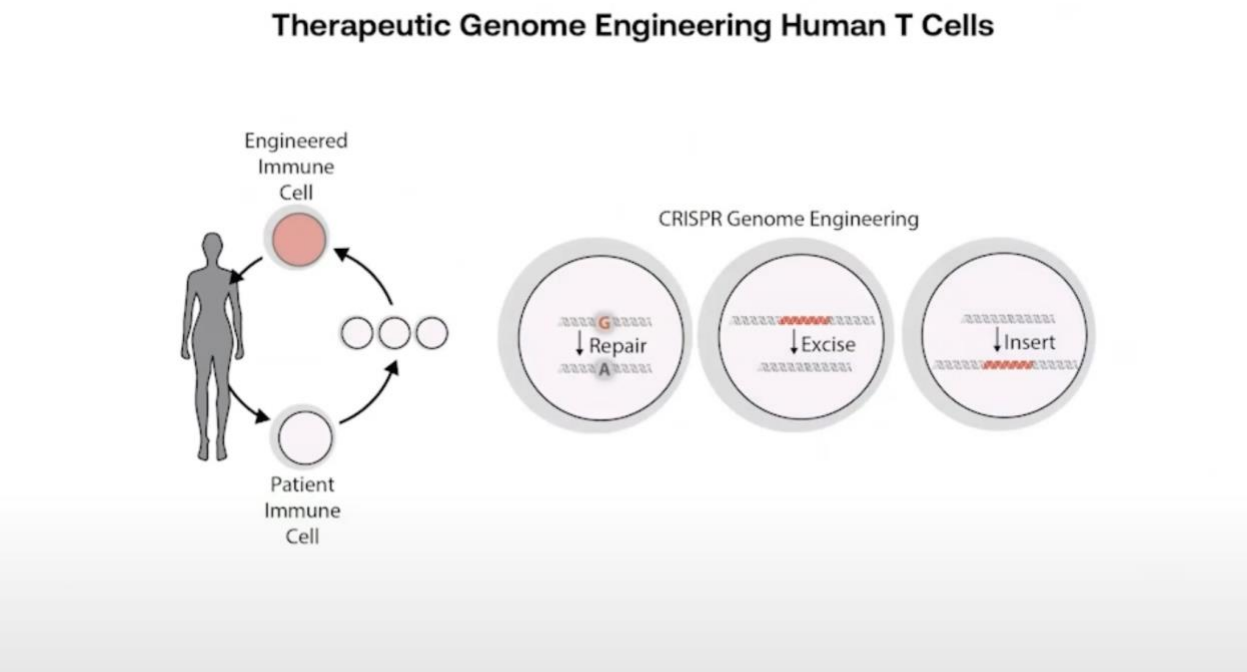
Clinical Gene Engineered Adoptive Cellular Therapy



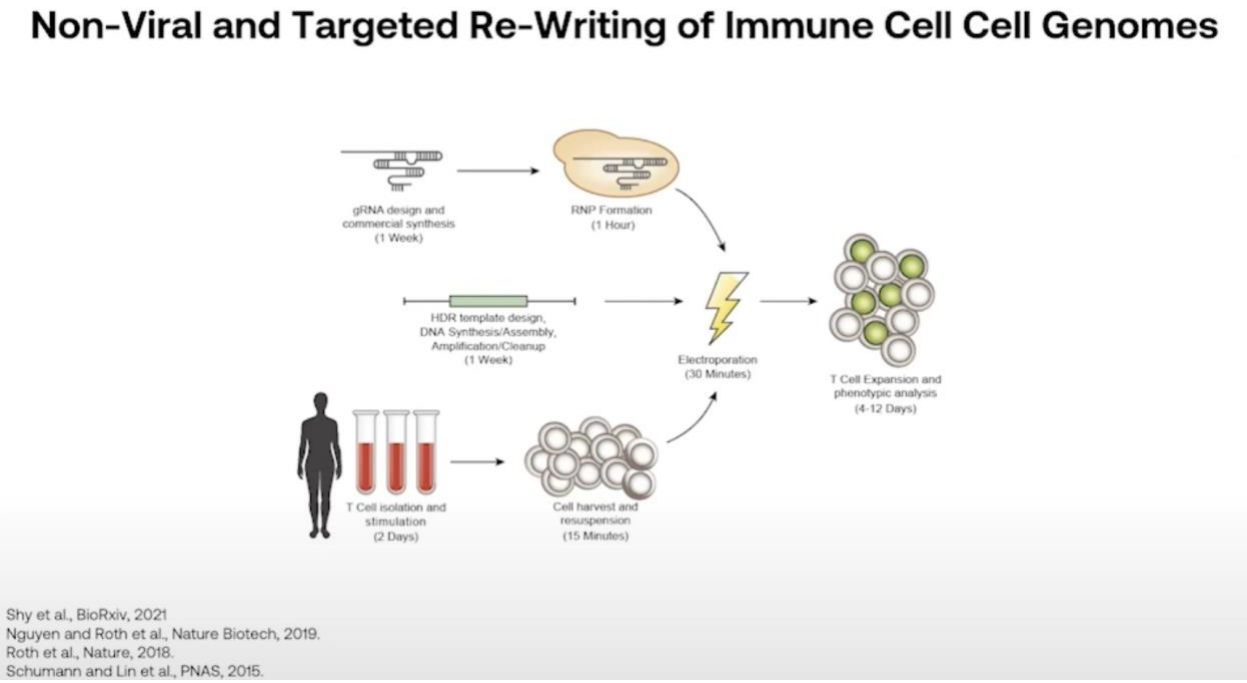
There are now several different responses for different blood cancers, but the goal is to create more powerful, durable, and lasting responses for patients with diseases that are more challenging, including solid tumors, where CAR T-Cell (still) have largely been ineffective and to other classes of human disease.
The current goal is to think about how to engineer a menu of cell therapies without only using viral vectors and take advantages of powerful new gene engineering technologies that allow us to pick very specific targeted sites in the genome and start rewriting DNA sequences.

To fix individual mutations in the case of rare disease, but also to think about what we can do to broadly reprogram cells by removing genes from the genome that are limiting the function of cells or by adding new DNA sequences to install a new gene program at a defined site in the genome of human immune cells.

Strictly Confidential 2023 Prepaire Labs Ltd



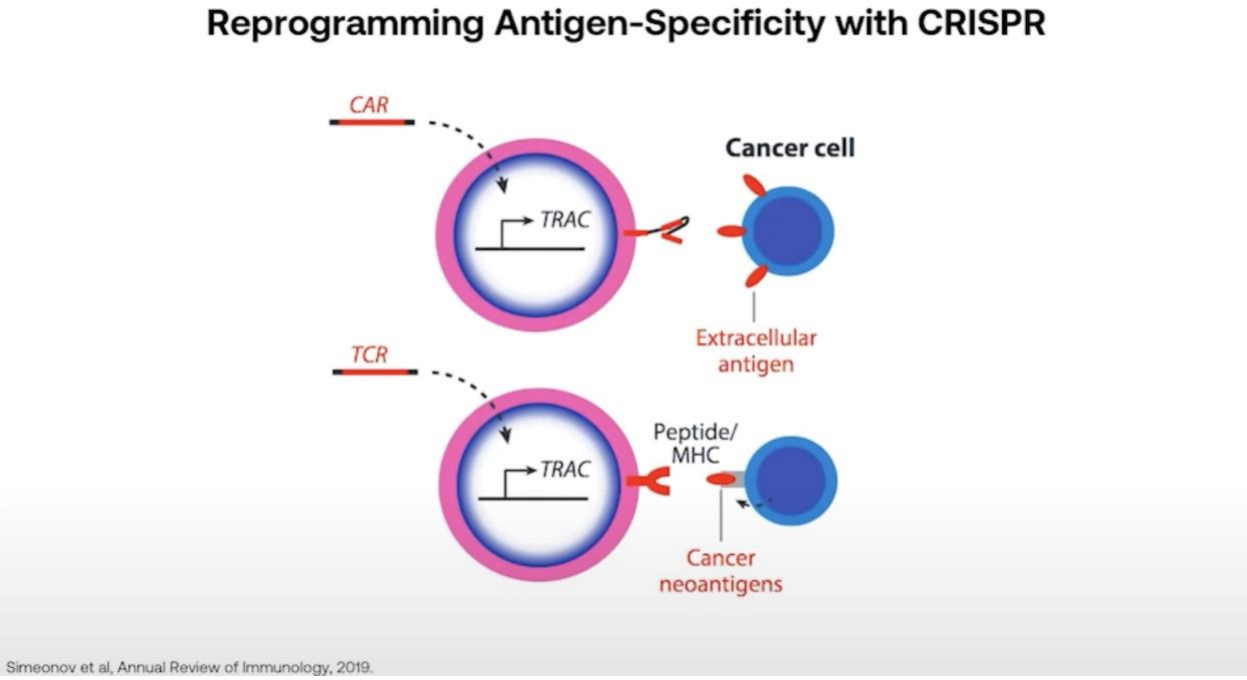
Over the past number of years, Dr Marson's lab have developed a technology platform that allows you to efficiently rewrite DNA sequences in a non-viral way at targeted sites in human immune cells.



This started several years ago through a collaboration with Jennifer Doudna and her CRISPR team. They started using recombinant Cas9 protein loaded in vitro with a guide RNA to make a ribonucleoprotein or RNP and found it that RNP can be electroporated into human T-cells to get very efficient gene disruption. Their lab also discovered that you could co-electroporate a non-viral piece of DNA that has homology arms on either side of the cut site to get very high efficiency knock-in sequences and pasting in a defined sequence at the defined site in the genome.

Strictly Confidential 2023 Prepaire Labs Ltd

They have now made several technological advances with new applications that extend the efficiency of this and with this is now a platform that's broadly applicable to several different cells in the immune system that can be cultured.

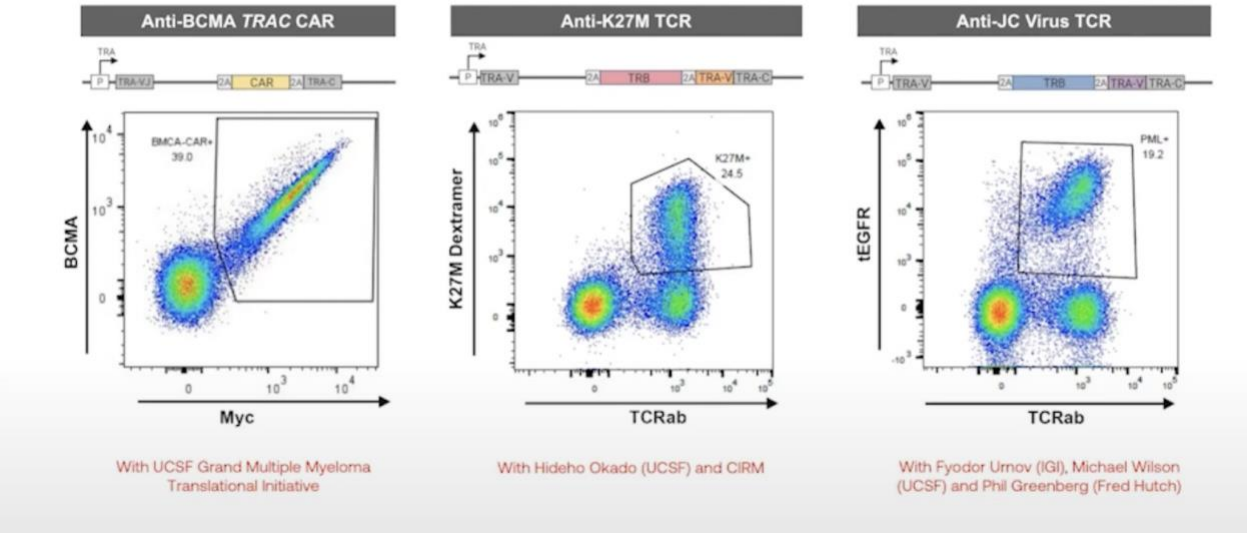


One of the most powerful methods would be to paste into a defined site, the antigen receptor and direct what a T-cell will recognize. This can either be done by pasting in a sequence for a CAR as described, or for by directly replacing the cells own T-cell receptor, which is the endogenous mechanism by which it recognizes an antigen.

There is an FDA approved BCMA CAR to treat multiple myeloma. The plan is to convert this into a non-viral method. The new methods clear 40% above knock-in of CARs with this non-viral process in a way that is compatible with clinical manufacturing. Dr Marson’s lab is working with the multiple myeloma translational initiative at UCSF and with the living therapeutics to advance towards a clinical trial for non-viral CAR T-cell. They are also working on a non-viral version of a TCR that recognizes an antigen found in certain types of pediatric brain cancer. And they are actively extending this beyond cancer.

Strictly Confidential 2023 Prepaire Labs Ltd

Towards Clinical Non-viral Knockin for CAR-T and TCR Therapies

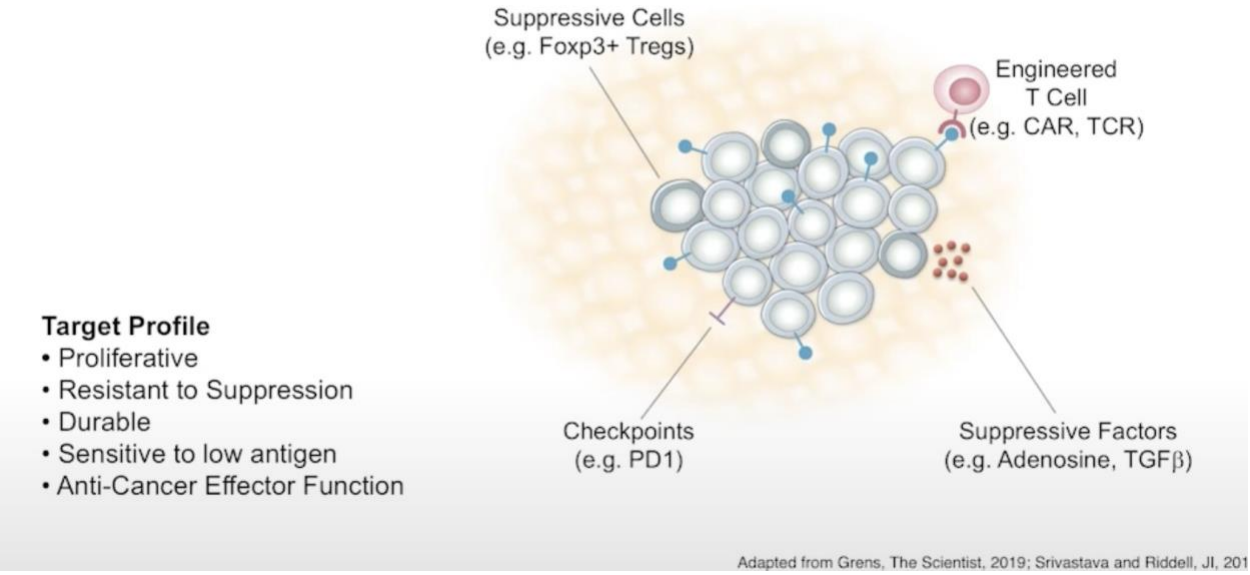


In addition to the above they are working on genetically engineered TCR cells that recognize viral peptides. The first step was dealing with sequences in the JC virus, which causes a horrible neuro infectious disease for certain immunocompromised individuals.
I think this is the tip of the iceberg of reprogramming antigen specificity, of different T-cells to go after cancer and infectious disease but also autoimmune disease.

The challenge making an antigen specific T Cell with CAR or TCR, is that when the cell encounters a solid tumor it will face several challenges that can prevent it from being effective in clearing the tumor. Once it gets into the tumor microenvironment, it will encounter the challenge of chronic antigen exposure and checkpoints, which can cause exhaustion and other modes of dysfunction. It will encounter suppressive cells and suppressive soluble factors.
All these need to be overcome.

Strictly Confidential 2023 Prepaire Labs Ltd

Overcoming the Challenges to Cellular Therapies For Solid Tumors



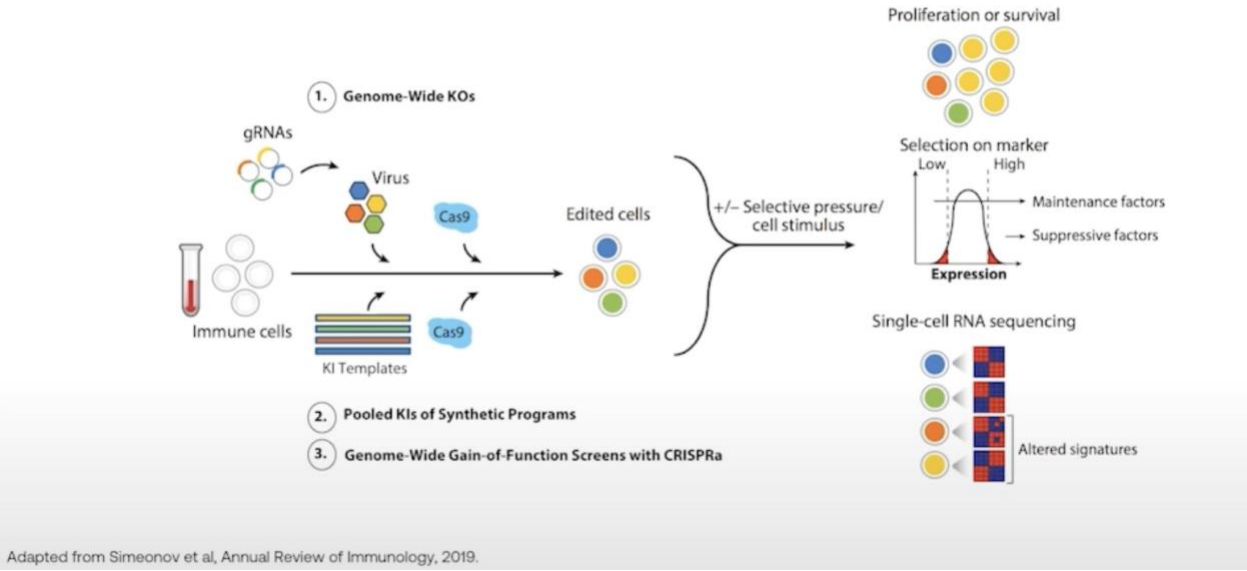
To design an idea cell-therapy, there needs to be an array of targeting designs which can guide us in being more effective.

- It should proliferate when it finds its target.
- It should be resistant to suppression.
- It should be durable.
- It should be able to sense low antigens.
- It should ultimately be effective at clearing tumors.

The idea is to create a development tool which explores all desirable target profile features. Introducing **FORWARD GENETICS**:

Strictly Confidential 2023 Prepaire Labs Ltd

Functional Genetics and Synthetic Biology in Human T Cells



Adapted from Simeonov et al, Annual Review of Immunology, 2019.

Forward genetics goes back at least as far as Drosophila studies, but we're not interested in applying this to a model organism. We'd like to use the unbiased power of forward genetics to discover genetic modifications in a primary human cell type.
That is the basis for cellular immunotherapy. And so, the idea of forward genetics is that you must take advantage of mutations that are introduced into a population. And then you pull out the individual cells in this case that have a phenotype of interest.

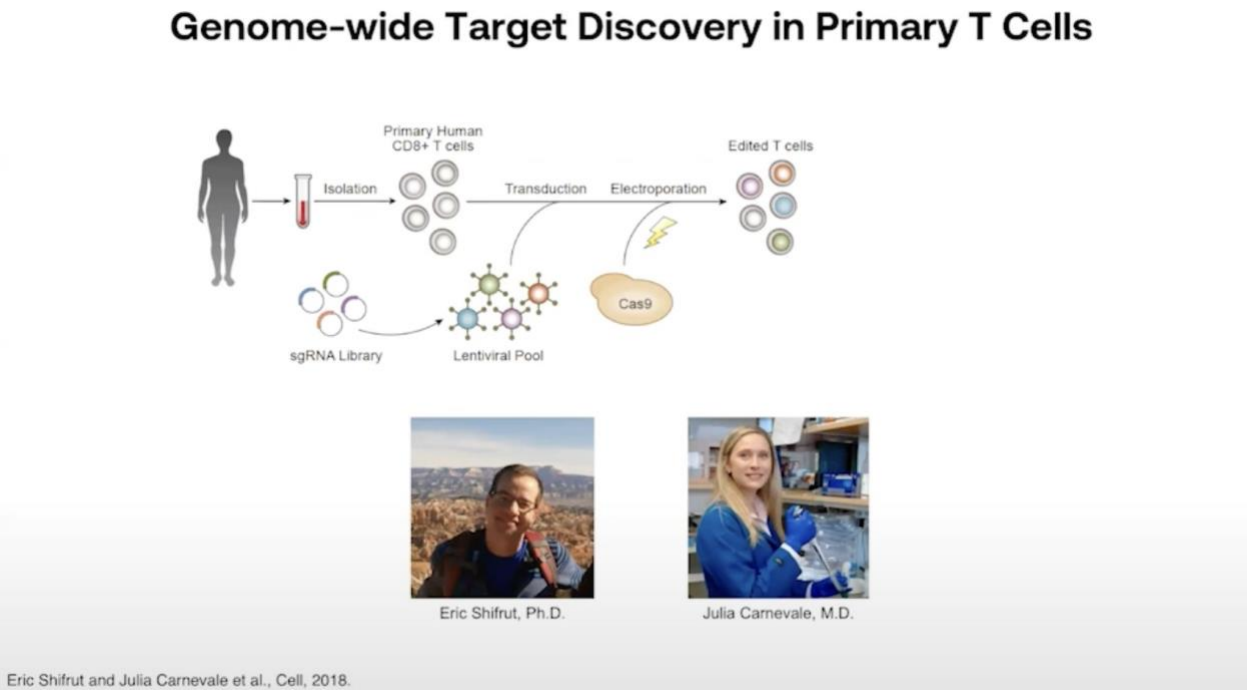
So how do we introduce large number of mutations into cells?
CRISPR has given us the basis to do this in extraordinarily powerful ways.
The solution lays in the ability to use a combination of CRISPR based functional genetic screens to pull out phenotypes of interest.
Marson and his lab are using genome-wide knockout screens with CRISPR.
They also moving beyond that into things that can be added into T-cells.
The result is a pooled knock-in to see if there are DNA cassettes that can be put in and then explore which one's drive phenotypes of interest. This also opens the door to conduct genome-wide Dayna function screens using CRISPR activation systems.

These technologies allow them to create a population of genetically modified cells that are effectively bar coded by different DNA sequences!!
And once they have a population of genetically modified cells, they can pull out the cells that have a phenotype of interest. The user can look for the modifications that make the cells more abundant in a successful situation, but also look for markers of interest.
You can couple this with single cell technologies.
So, you can use the power of genome-wide CRISPR screens, which have been widely used in cell lines and animal models to actually do this in primary human T-cells.

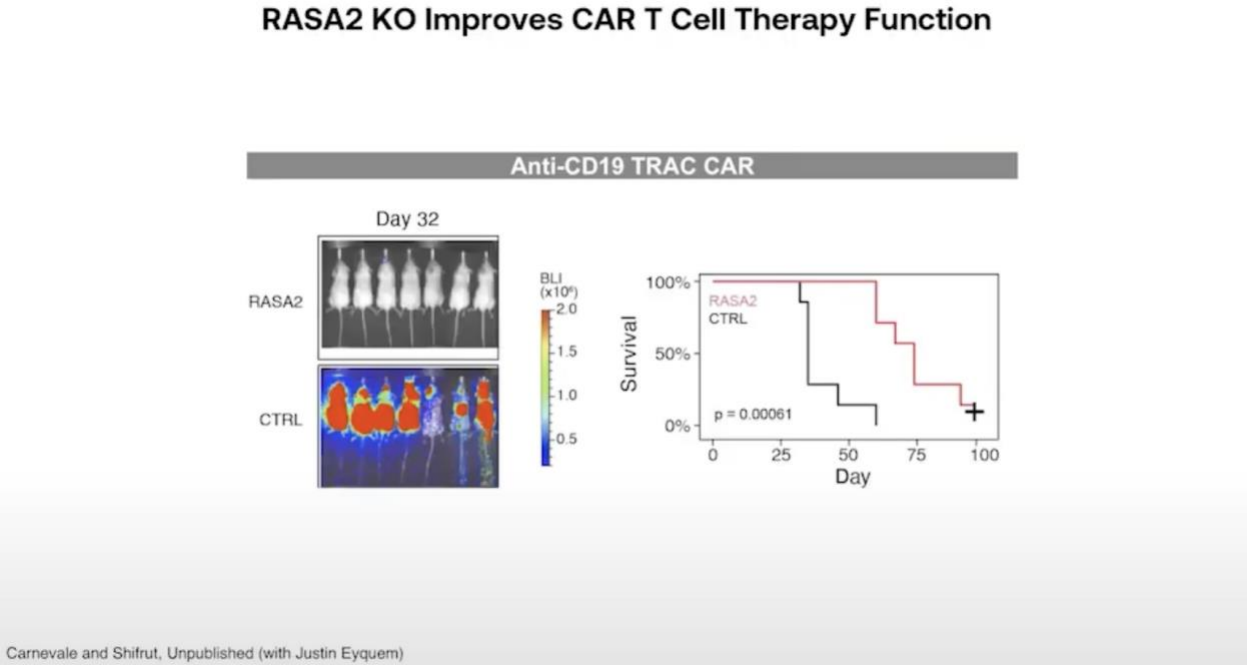
Marsons lab have developed a system to do a genome-wide screen where you put libraries of antiviral vectors into cells that carry the guide RNAs. And then because they had trouble getting Cas9 in with virus, they electroporated the Cas9 protein to do very efficient gene editing. And then when you had this population of cells, you could screen for any phenotype of interest and they've

Strictly Confidential 2023 Prepaire Labs Ltd

systematically dissected gene modifications that make cells proliferate more when they get restimulated.

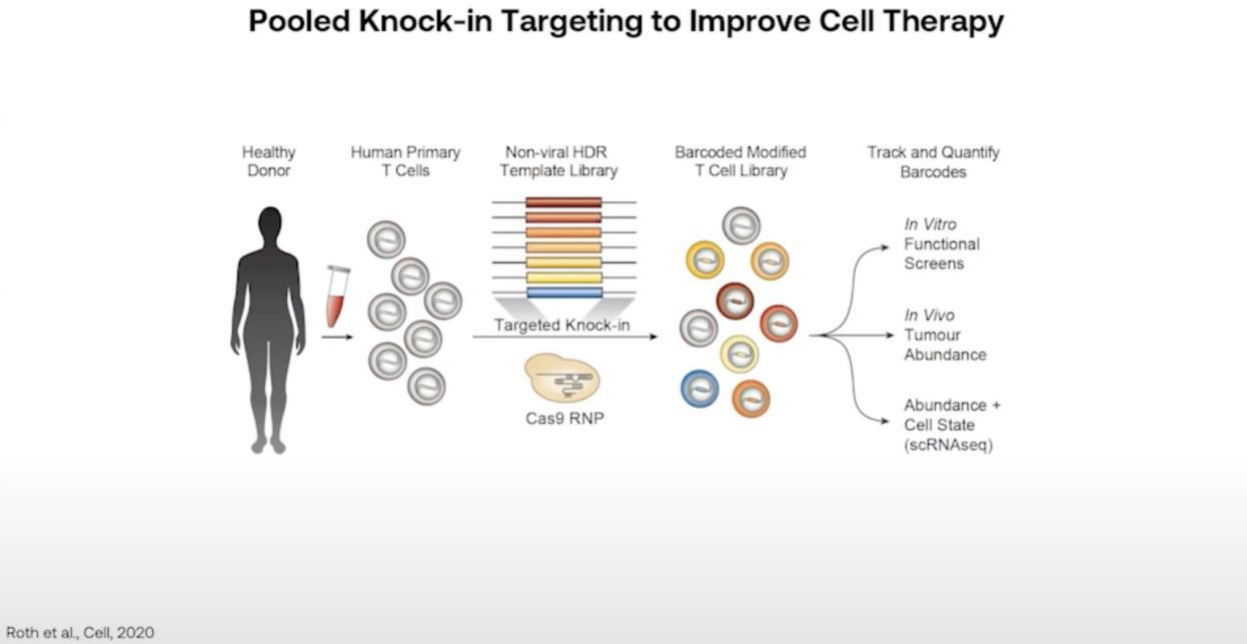


This got me excited about the power to do high throughput screens, to discover gene modifications that will make cell therapies more powerful. And now we're really thinking about it, can we go beyond what we can take out of a gene out of a cell?
What can we add into a cell that will enhance its function?



The Prepaire™ user objective should be to introduce antigen receptors in gene programs, then aiming into targeted sites with CRISPR. Using TCR or a CAR plus an extra gene where

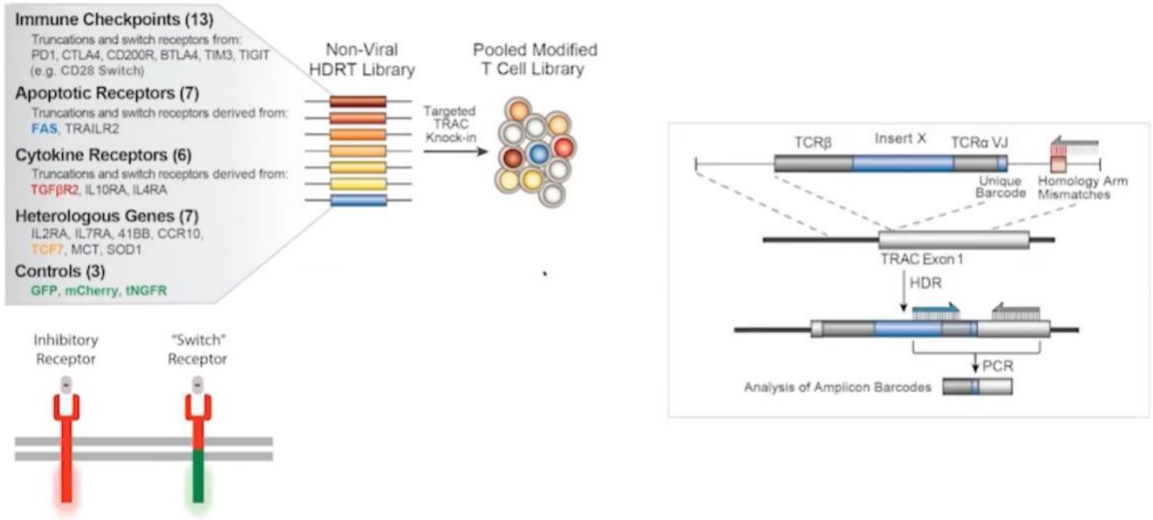
that extra gene may have a beneficial function for promoting a cell state of interest in cells, in cell therapies. And rather than guess, you create a computer model (test) as an unbiased way to make many different DNA constructs that might have beneficial effects and knock them all into a cell of interest and then effectively race them against each other in vitro or in vivo, or analyze them all with single cell sequencing to see which of those programs is most productive in moving cells to a state that is compatible with effective cell therapies.



The Marson lab has already started making a library of candidate constructs, either genes that could be overexpressed in the TCR alpha locus and make constituently expressed in T-cells, or even thinking about synthetic receptors that don't exist in nature, or so-called switch receptors. This is where you make an artificial fusion what would otherwise be an inhibitory receptor and keep the extracellular domain that recognizes this inhibitory ligand and start fusing it to an activating intracellular domain.

Strictly Confidential 2023 Prepaire Labs Ltd

Gene Programs to Enhance T Cell Function



Roth et al., Cell, 2020

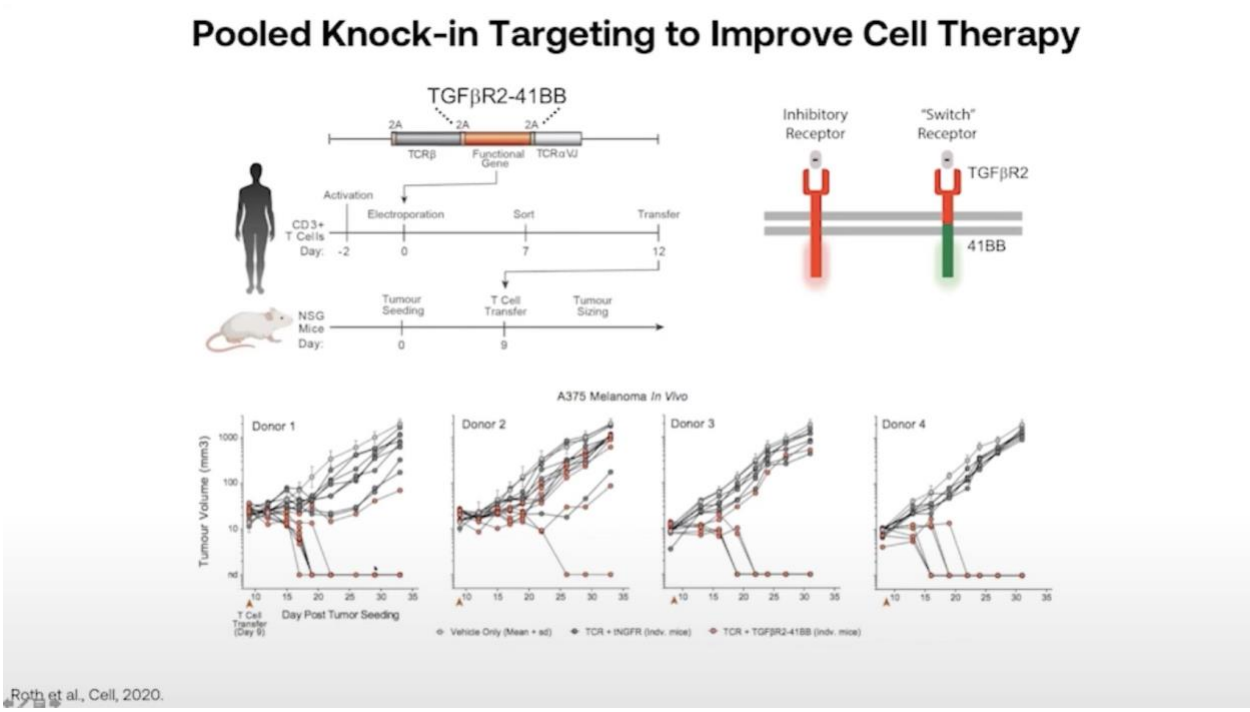
This switch receptor should then respond to something that would normally suppress T-cells in their microenvironment, but instead makes them strongly activated. The lab designed a candidate library of 36 different members, including controls natural genes and synthetic receptors. Then knocked them into the TCR locus along with a defined antigen specificity.

Each cell got a different knock-in and they could track which knock-in went into which cell in its population based on a barcode that was put in along with them.

It was so powerful that they conducted an in vivo study to see which of these knock-ins constructs made the cells accumulate more in a tumor microenvironment.

These are human immune cells knocked-in, and each one got the same T-cell receptor, but with a different gene edit along with it.

Strictly Confidential 2023 Prepaire Labs Ltd



This is the T-cell receptor we used recognize an NY-ESO cancer antigen.

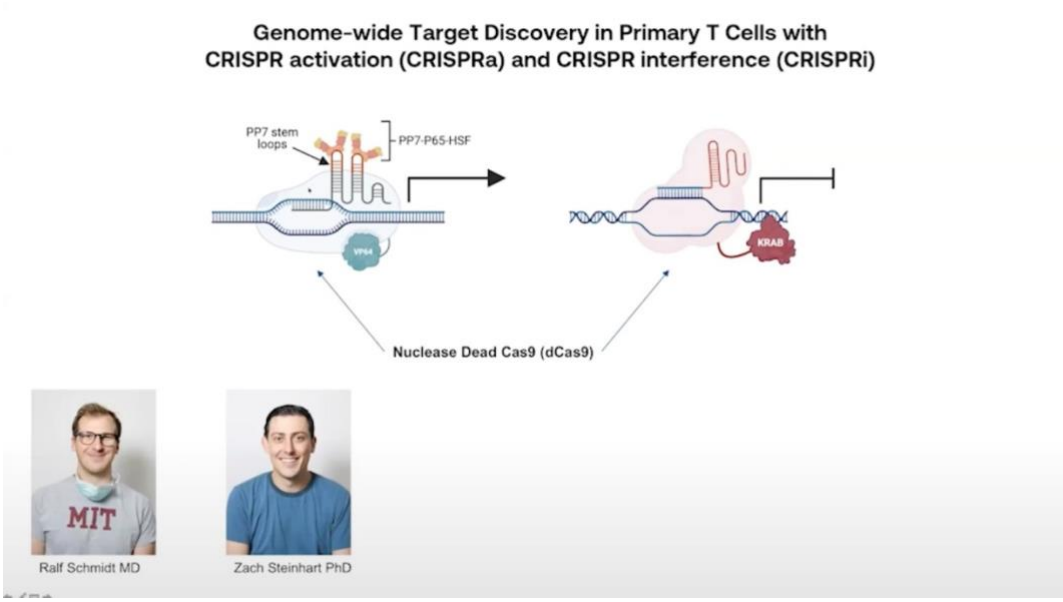
After inserting an NY-ESO positive tumor into an immune deficient mouse, they compared the input population of cells to the cells that infiltrated them, accumulating in the tumor and then examined which barcodes led to preferential accumulation of the tumor and started finding genes, including these artificial receptors that manipulate TGF beta sequences that lead to preferential accumulation.

These TGF beta switch receptors that couple and act on these inhibitory TGF beta receptor to activate intracellular domains seem to lead to tumor microenvironment accumulation.
Testing that one at a time looking at the TGF beta 4-1BB construct that had been generated, then placing it with an antigen receptor into a mouse.
The results were really striking. Compared to the control mice that don't get this hormone BB receptor with the TCR, the TGF beta 41 BB switch receptor leads to complete responses in a number of these mice at this dose, which was not otherwise achievable at this dose.

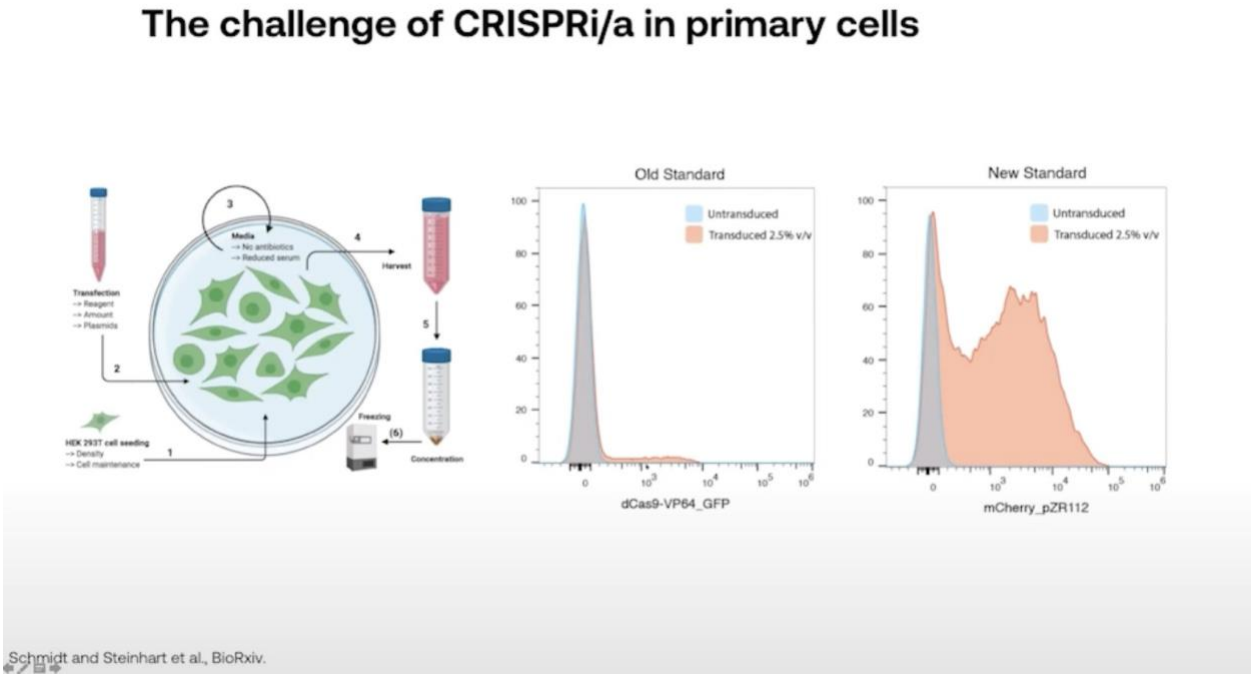
This opens the idea of how to scale this, and that could be where Prepaire GPU/CPU horsepower comes into the picture.
How do we discover even more rapidly what can be overexpressed in the cell to make it more effective at doing what we want to do in several different disease contexts?
You would like to be able to discover systematically every gene that could be overexpressed either on its own or put into a synthetic circuit.
There is a very powerful technology to do genome-wide screens that would allow for data function. But this has only been done in cell lines.

Introducing CRISPR activation.

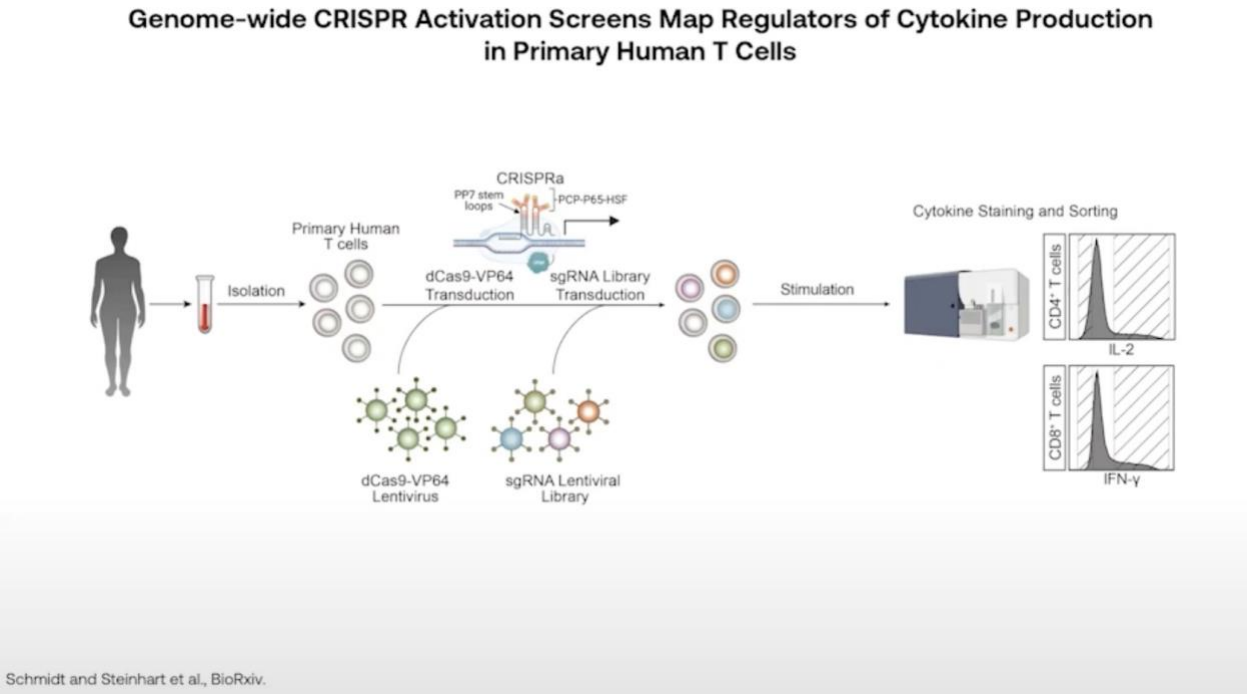
Strictly Confidential 2023 Prepaire Labs Ltd



This approach uses a dead Cas9 which is neither doing cutting or gene editing.
The Cas9 is only being used to recruit transcriptional activators.
Several different transcriptional activators that are fused to the dead Cas9 itself and to the guide RNA to be, turn on the genes of interest. There's another version of this called CRISPR interference where the dead Cas9 brings a transcriptional repressor.
The idea is to use these modalities to turn genes on and off. But the challenge has been that unlike Cas9, where we could deliver a transient pulse of Cas9 protein and get permanent genetic modifications, these technologies need to be sustainably introduced into T-cells, which means that really to use them in practice, we need to be able to get stable transduction with viral vectors. That has been a big challenge because these are big constructs, and the efficiency has been low.



The approach is to use machine models to systematically start testing several different variables to see what could be optimized to get higher efficiency. We could start introducing CRISPR activation system into primary human cells.
But this just gives you a flavor of putting in this SAM system that was developed at the Broad into human T-cells. Now suddenly, compared to the control cells with this control guide RNA, which don't express these receptors measured by flow cytometry.
You could get robust induction across multiple different guide RNAs using this CRISPR activation system.
Using CRISPR activation to introduce, to activate a different gene from every gene in the genome, in a different cell by using a combination of this CRISPR activation virus along with a separate virus that puts in the guide RNAs and introduce them both into cells.
That way you can create genome-wide libraries that introduce gain of function perturbations in the population of primary human cells. Once you have that they could, in theory, test any different phenotype of interest to see the gain of function.



But what are the genes in the genome that can be tuned to change how the cells will respond to stimulation?
We know that as cells respond to stimulation, CD4's and CD8's, depending on how they're stimulated and the state of the cells, they will turn on different cytokines, which will have a big effect on their effector functions.

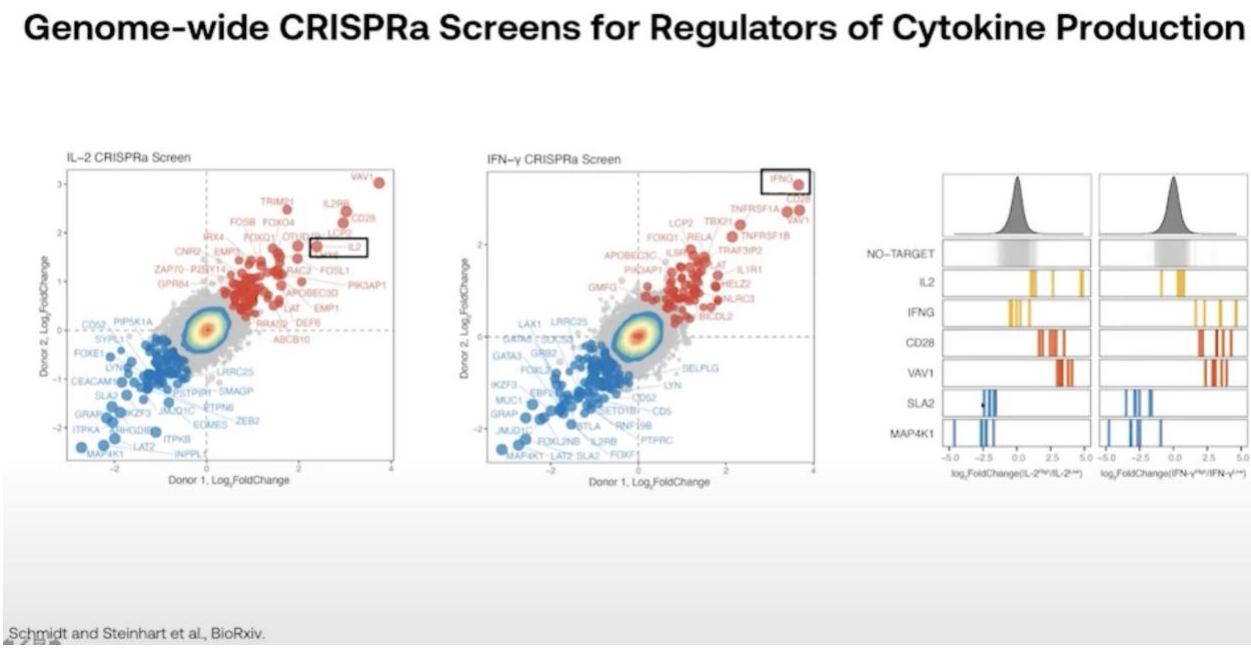
By looking at CD4 cells to understand the genes that could affect the extent to which IL2 is produced and in CD8's, what are all these genes that would affect the extent to which interferon gamma is produced? So now they're looking at endogenous cytokine production.
These are not reporters. They're doing intracellular staining and sorting for cells that have different levels of the cytokines produced after perturbations.

Here are the results of Geno-wide screens for all the gain of functions that modulates levels of IL2 and interferon gamma production: *These were done in two separate human blood donors. So even though these were from different individual humans, the results were remarkably reproducible across these genome wide experiments.*

Strictly Confidential 2023 Prepaire Labs Ltd

- This linear relationship shown for both screens and the results make sense.
- Among the top hits, if you activate IL2, you get more IL2 production.
- If you get more, you activate interferon gamma directly, that's the top hit for interferon gamma high cells.

It was also uncovered many other regulators that when you overexpress them lead to more cytokine production or less, and these were reproducible not only between human blood donors but were strongly reproducible among multiple different guide RNAs suggesting that these are actual effects.



Here you can see IL2 is a strong hit in the IL2 interferon gamma, and then several hits that when they're overexpressed as expected CD28 or Vav strongly affected cytokine production, as well as overexpression of negative regulators, that would dampen this cytokine production.

Next was screening with CRISPR interference. And again, the results were highly reproducible across human blood donors. And now the hits come up IL2 over interference, lowers IL2 as expected, and interferon gamma lowers the interferon gamma.
And they started to really see complimentary maps of genes that can be up-regulated or down-regulated to tune how what cytokines will be produced as CD cells are responding. CRISPR activation and CRISPR interference are really providing useful and complimentary maps.

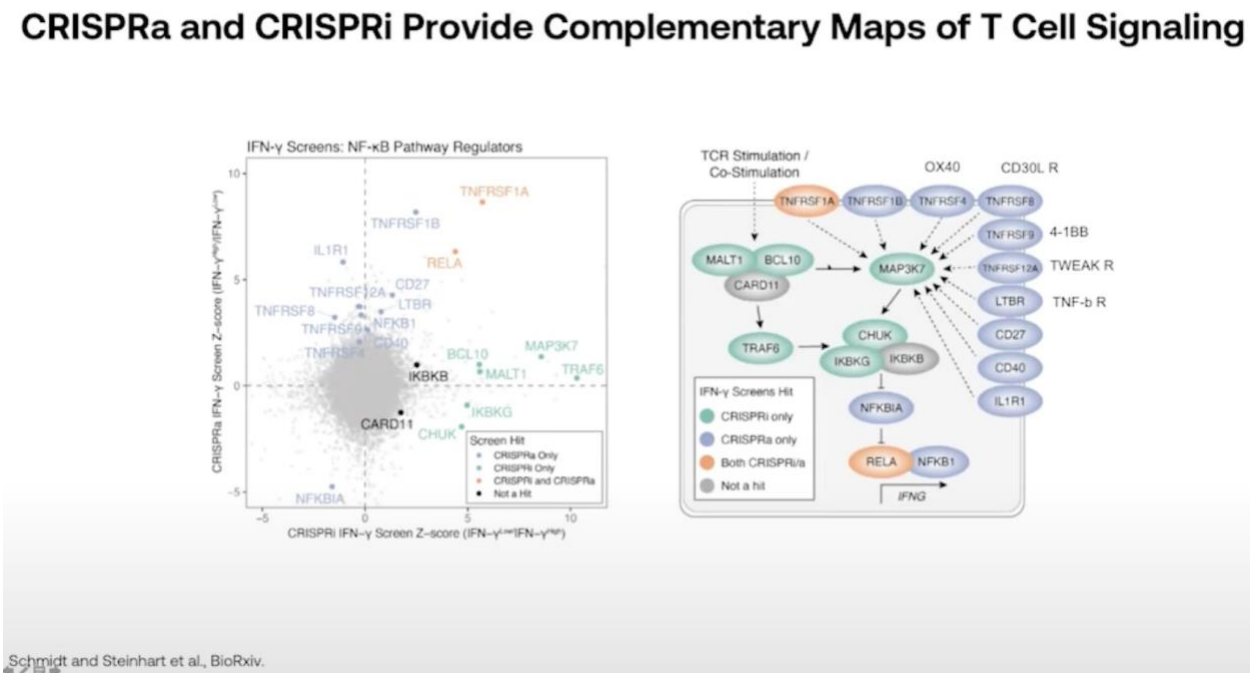
If you look at collectively all the hits from the CRISPR interference screens, the hits as you'd expect are, tend to be well expressed in the cells that we're screening, because you will only find a phenotype with interference if that cell is if that gene is already on. In contrast, that is not true for CRISPR activation.

Many of the hits, they hit the range of expressions, much bigger. And many of these are expressed at very low levels in the cells that were screened.

You can detect gene function independent of the context where these genes are normally activated, normally used. And that allows us to start putting together these things, to see, to look at whole

pathways, where we can see what comes up as a hit with CRISPR interference or CRISPR activation. If we look at the NF kappa B hits across the interferon gamma screen, for example, we see that a few hits only come up in CRISPR activation and others only come up in CRISPR interference.

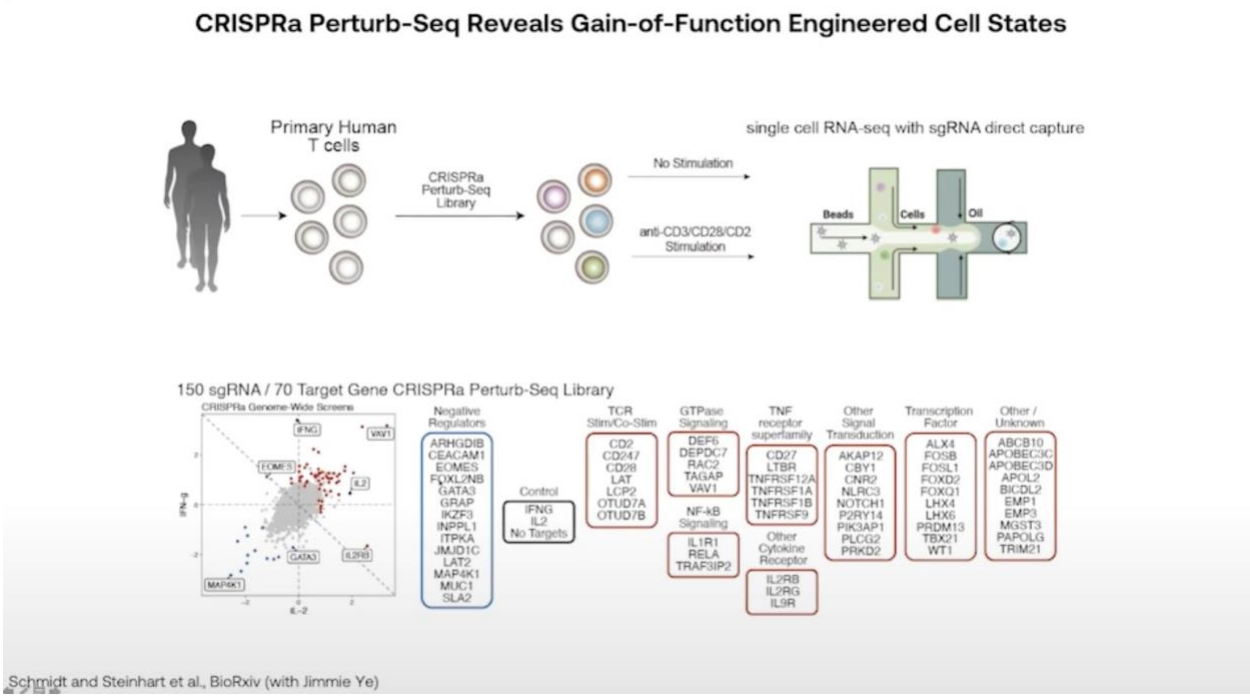
Several intracellular components, for example, they seem to be critical and non-redundant and expressed well at basal levels in the cells. And these come up as CRISPR interference hits, in getting rid of these, blocks the ability of cells to make interferon gamma.
In contrast, there's a whole number of cell surface receptors, including many of these **super family receptors** that are only come up in CRISPR activation.
Many of these may or may not be expressed at sufficient levels under these cells, but we can pick up that these are sufficient to turn on interferon gamma and independent of the context where they may otherwise be used. And this is not just an arbitrary list.



These include many things that point us to potential immunotherapy targets, which you'll recognize here, including OX40 and 4-1BB and CD27 CD47, which are being actively developed for a number of different types of immunotherapies.
So now we're discovering receptors that can be activated to drive cells to states of interest.
CRISPR activation can be coupled to unbiased single cell measurement at cell state.
To really tell us what is regulating an individual cytokine of interest, and what's the overall state that we're promoting by putting in this gene of interest into a T-cell. By adapting this CRISPR activation perturbation system in primary human T-cells it became compatible with 10X droplet based single cell sequencing.

The result is that each one of these perturbed cells, we can not only look at the overall transcription, but we can look at, we can measure which guide RNA is actually in that cell.
So, we can ask when a gene of interest is overexpressed, what's the effect on overall cell state?
They picked out 70 different genes, including controls and hits from their screens, things that generally promoted cytokine production, as well as things that dampens cytokine production when they were overexpressed. And they did a CRISPR Perturb-seq node, in primary human T-cells to look at with a single cell transcriptome read out what is the effect of overexpressing each of these targets or controls.

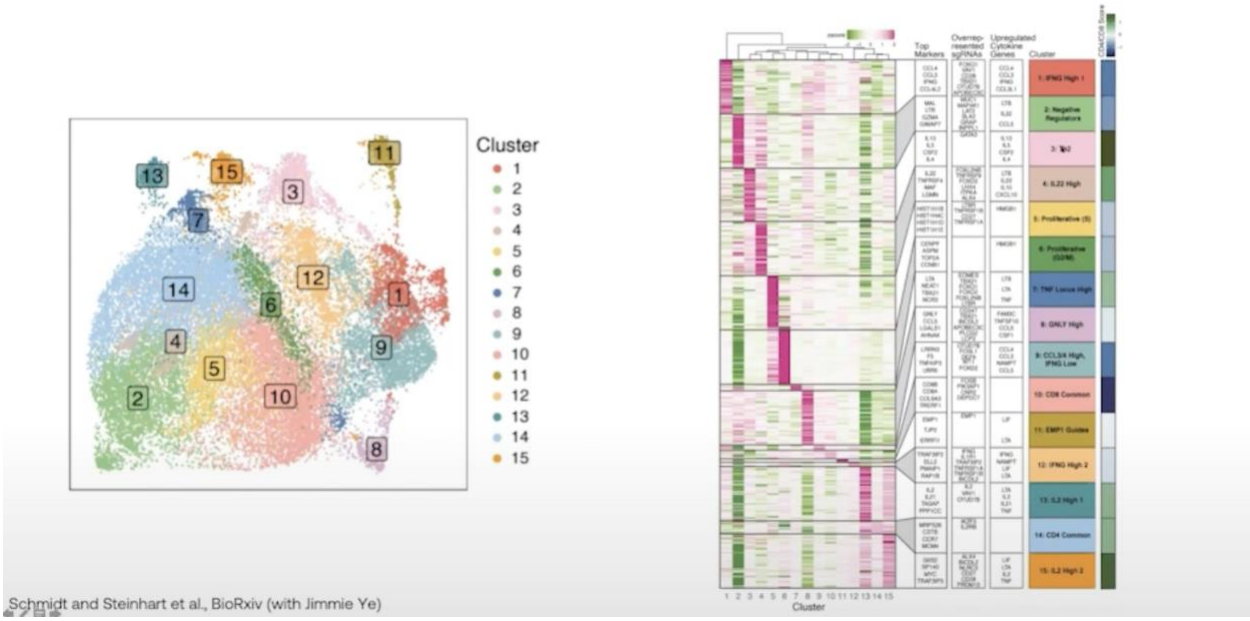
Strictly Confidential 2023 Prepaire Labs Ltd



You can see very high degree of mixing of these cell states, independent of blood donor.
They were able in a both population of stimulated T-cells to resolve CD4 very well versus CD8 cells.
You can separate CD4 and CD8 cells and ask individually each of these perturbations, what do they have, what effects do they have on both CD4 and CD8 cells?
And then we're able to start looking into this into the cells in this UMAP and say, what cytokines are they producing?

If we look at this heat map or which cells are producing interferon gamma, you see that they tend to cluster in a few different places here and here. IL2 was captured less well at the single cell level, but you can still see IL2 cells that cluster together. And then with this map of all the cell states in the population, what we really want to know is which perturbation is driving cells selectively to different states that can be measured on this map.
The control cells are evenly distributed across the cell state map as you'd expect.

Strictly Confidential 2023 Prepaire Labs Ltd



And then if we look at the cells with all the perturbations, they start to cluster in different states. And we can look at one perturbation at a time in this pool to experiment and start to really make sense of it. So overexpressing MAP4K1 which was a negative regulator, dampened activation, and move cells to this state, they corresponded with overall low levels of cytokine production. And then we were able to see individual regulators.

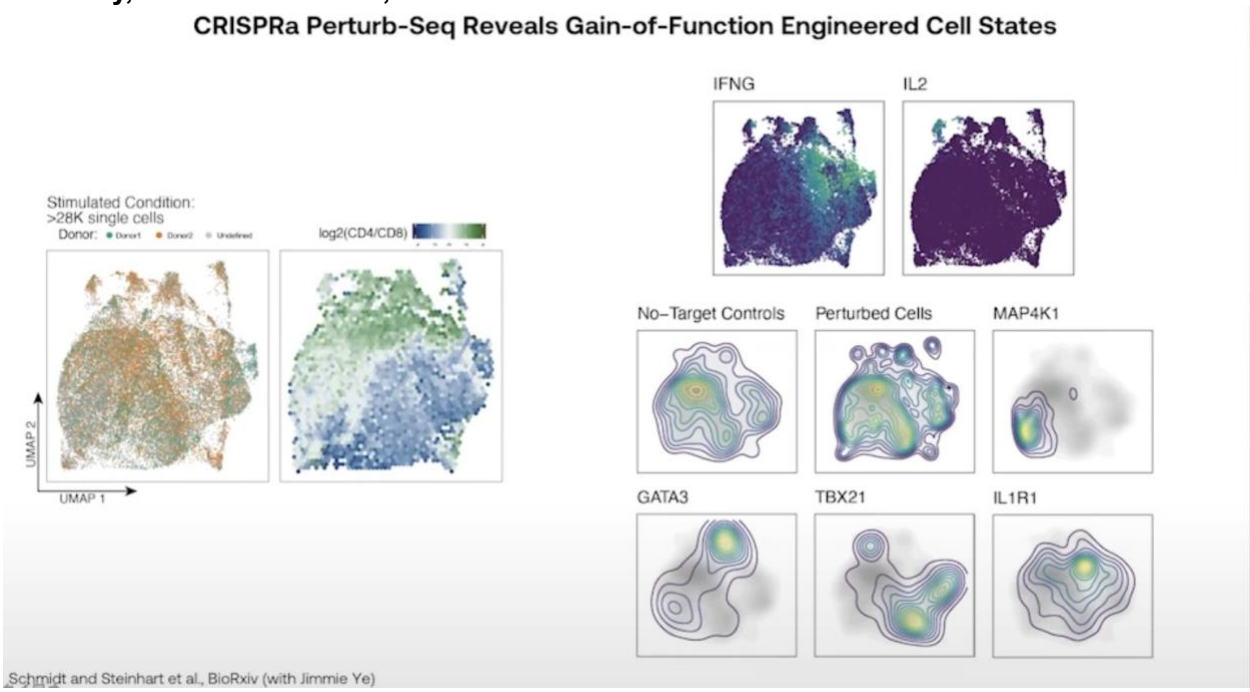
So here were a couple of canonical regulators overexpressing GATA3 for example, moves the cells to a strong TH2 state T-bet move cell to a TH1 state or a interferon gamma producing, CD8 state. And then we could also find this for less well-characterized regulators.
Met overall single cell measurements of cell states and for the aficionados, what the lab accomplished was to make an unbiased, relatively unbiased map of naming the different cell states that come up in this map and systematically map, these different clusters to which guide RNAs are moving them to these different cell states.
Among these different 15 clusters, we were able to name many of them and they become recognizable. For example, proliferative cells or TH1 cells, TH2 cells and ask which guide RNAs are found in each of these clusters. And you could see that here that there are certain guide RNAs that are moving cells to define cell states.

From finding the genome-wide level of regulators, cytokine production, to be able to pair individual perturbations up with very detailed maps of cell state.
The opportunity for Prepaire is to help create a tableau that defines the states that will best characterize the most effective cell therapies for cancer and for other diseases and map the gain of function and loss of function perturbations to directly program cells into the states that we will make them best able to treat patients for several different diseases.

These discoveries have enabled a suite of forward genetic technologies that allow us to perturb genetic elements throughout the genome in a function loss of function.
Knock-in everything about coding sequences, but now actively moving this beyond into non-coding sequences.

Strictly Confidential 2023 Prepaire Labs Ltd

This is applicable across different cell types. This is going to tell us and is already telling us fundamental ways that the genome programs, the behavior of human cells. It's pointing to new drug targets, but it's also giving us the basic map of constituent elements that we can use to start programming cell therapies and everything we learn can has a direct path to being the basis for the roadmap to design next generation cell therapies for cancer, auto-immunity, infectious disease, and others.



Strictly Confidential 2023 Prepaire Labs Ltd