

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Abschlussbericht

Zuwendungsempfänger: Max-Planck Gesellschaft

Projektleiter: Prof. Dr. Michael Ziller

Projekttitel: MERGE: Model Exchange for Regulatory Genomics

MERGE: Modellaustausch für die regulatorische Genomik

Förderkennzeichen: **031L0174B**

Laufzeit des Projektes: 48 Monate + 12 Monate Verlängerung

Berichtszeitraum: 01.04.2019 - 31.03.2024

Kontaktperson: Prof. Dr. Michael Ziller
Universität Münster
Busso-Peus-Str. 10
48149 Münster
ziller@uni-muenster.de
+49 (0)251 / 83- 34422

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

I. Kurze Darstellung

1. Aufgabenstellung

Computergestützte Modelle der regulatorischen Genomik sind die Grundlage für die Interpretation personalisierter Genomvariationen, die die Mechanismen krankheitserregender genetischer Varianten oder somatische Mutationen bei Krebs aufklären. Trotz des erheblichen Potenzials dieses Feldes fehlt es jedoch an kohärenter Software, Pipelines und Schnittstellen, um Methoden für die regulatorische Genomik effizient einsetzen zu können.

2. Voraussetzungen

Um dies anzugehen, werden wir auf einer Prototyp-Implementierung von Kipoi aufbauen, ein Modell-Zoo und Repository für regulatorische Genomik. In diesem Projekt werden wir Kipoi erweitern, um die Interpretation von persönlichen Genomen und Modellen zu unterstützen. Darüber hinaus werden wir Modellen wichtiger Genregulationsprozesse entwickeln, die auf Beobachtungs-Omics-Datensätzen und Hochdurchsatz-Perturbationstests basieren und passgenaue gute Praktiken für die kohärente Kombination von Modellen liefern, welche auf heterogenen Omics-Daten trainiert sind. Schließlich werden wir prototypische Pipelines für die genetische Varianteninterpretation bei seltenen und häufigen Erkrankungen sowie in Tumorgenomen etablieren und anwenden.

Um diese Ziele erfolgreich zu erreichen, haben wir ein Konsortium aus drei Forschern mit komplementären Fachkenntnissen gebildet. Die Expertise der Gruppe von Julien Gagneur besteht in statistischer Modellierung von Transkriptomdaten und transkriptionellen Prozessen, insbesondere von RNA-Verarbeitung, Epigenom und Systemgenetik. Dr. Oliver Stegle ist Abteilungsleiter am Deutschen Krebsforschungszentrum in Heidelberg und arbeitet an der Schnittstelle von maschinellem Lernen, statistischem Rechnen und den Lebenswissenschaften. Er hat bahnbrechende Computermethoden für Molekulargenetik, Einzelzellgenomik und für die Integration von heterogenen molekularen Datensätzen und klinischen Kovarianzen entwickelt. Dr. Ziller verfügt über umfangreiche Erfahrung auf dem Gebiet der funktionalen Charakterisierung nicht-kodierender genetischer Elemente sowie der Reprogrammierung somatischer Zellen zu pluripotenten Stammzellen und der *in vitro* -Differenzierung entlang der neuralen Linie. Insbesondere hat der Antragsteller erfolgreich verschiedene *in vitro* -Differenzierungssysteme entwickelt und implementiert, etwa die Differenzierung von humanen pluripotenten Stammzellen zu Vorläuferpopulationen der drei embryonalen Keimblätter (Ectoderm, Mesoderm, Endoderm) und der Differenzierung in eine Reihe unterschiedlicher neuraler Vorläufer- und neuronaler Zellen.

3. Planung und Ablauf des Vorhabens

MERGE ist in 5 Arbeitspakete gegliedert, um diese Anforderungen zu erfüllen (Abb. 1):

AP1. Kipoi model zoo Infrastruktur (Leitung: Stegle): Erweiterung von Kipoi, um seine Funktionalität und Anwendbarkeit auf große genomische Datensätze zu verbessern und sein statistisches Grundgerüst zu erweitern.

AP2: Zusammengesetzte Modelle auf der Grundlage natürlicher Variationsassays (Leitung: Gagneur): Entwicklung neuartiger Modelle, die einzelne bestehende Modelle wie Bausteine kombinieren, um Schlüsselaspekte, des zentralen Dogmas der Molekularbiologie, nämlich Transkription, Spleißen und Translation, zu modellieren.

AP3: Nutzung von Hochdurchsatz-Störungsdaten zur Konstruktion von DNA-Sequenzaktivitätsmodellen für die personalisierte Genominterpretation (Leitung: Ziller): Aufbau und Integration neuartiger Algorithmen auf der Grundlage funktioneller Genomdaten zur Verbesserung der Identifizierung kausaler Beziehungen zwischen einzelnen genetischen Varianten und funktionellen molekularen Folgen.

AP4: Variant effect predictor using Kipoi composite models (Lead: all): Integration von Kipoi-Vorhersagen unter Verwendung bestehender Modelle und der in WP2 und WP3 entwickelten Modelle, um eine verbesserte Variantenpriorisierung sowie personalisierte Vorhersagen zu ermöglichen.

AP5 Workshop über computergestützte regulatorische Genomik, um die in MERGE entwickelten Technologien zu verbreiten.

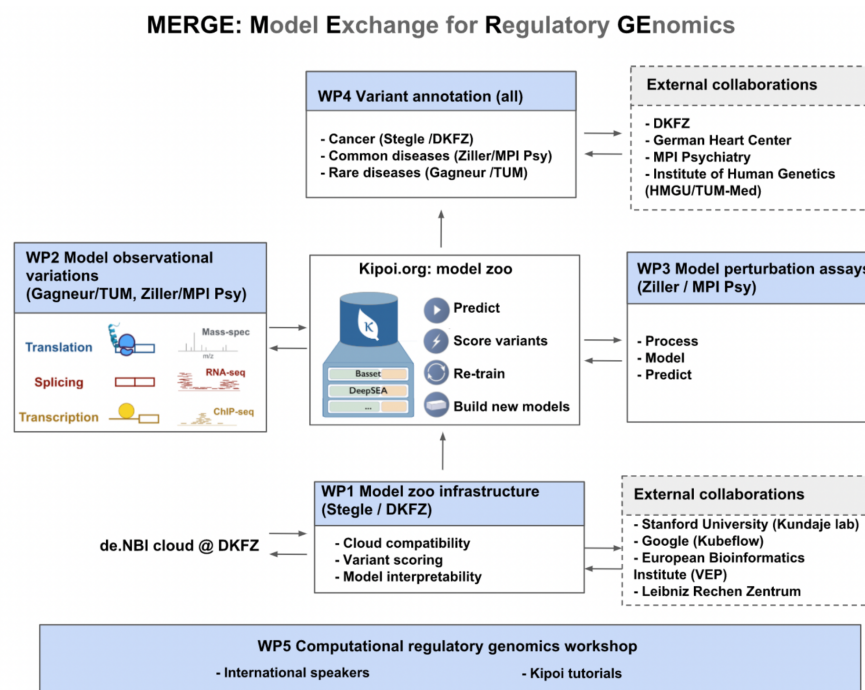


Abbildung 1 | Überblick des MERGE Projekts

Insgesamt wurden die Ziele des Projekts erreicht. Im Rahmen von WP1 haben die Gruppen von Oliver Stegle und Julien Gagneur erfolgreich eine skalierbare Software zur funktionalen Annotation von Varianten mit Hilfe von Deep-Learning-Modellen entwickelt und diese auf die gesamten 500.000 Individuen der UK Biobank angewandt (Clarke et al., 2023).

Im Rahmen von WP2 wurde von der Gruppe Gagneur sequenzbasierte Modelle für verschiedene Schritte der Genregulierung entwickelt, darunter Transkription und RNA-Abbau (Hözlwimmer et al., 2023), RNA-Spleißen (Cheng et al., 2021; Wagner et al., 2023), Sequenzdeterminanten der Translation (Tomaz da Silva et al., 2024) untersucht und die Methodik großer Sprachmodelle auf genomische Sequenzen großer Genomdatenbanken mit speziesspezifischem DNA-Sprachmodellen (LM) portiert, um konservierte regulatorische Elemente zu erfassen (Karollus et al., 2024). Wir haben außerdem ein Benchmark eines modernen sequenzbasierten Deep-Learning-Modells der Genexpression durchgeführt (Karollus et al., 2023) und ein verbessertes Modell der Enhancer-Promoter-Regulation entwickelt (Hecker et al., 2023). Schließlich hat das Konsortium erste Schritte unternommen, um die Interpretation von DNA LMs zu erforschen (Silva et al., 2024).

In WP3 wurden Methoden zur sequenzbasierten Vorhersage des funktionalen Effekts häufiger Varianten auf die Genexpression entwickelt. Zu diesem Zweck wurden verschiedene, aufeinander aufbauende Deep Learning Modelle etabliert, die unterschiedliche Arten von experimentell bestimmten biologischen Annotationen des Genoms (Epigenetische Informationen) bzw. Hochdurchsatz-Assays (MPRA Datensätze) nutzen, um den Effekt einer Single Nucleotide Variant (SNP) auf die Genexpression vorherzusagen.

Im Rahmen von WP4 wurde eine neue Methode namens DeepRVAT entwickelt, die die Annotationen seltener Varianten mit Hilfe von Deep Set Modellen integriert. Diese Methode ermöglicht verbesserte Assoziationstests für seltene Varianten und die Identifizierung von Personen mit hohem genetischem Risiko (Clarke et al., 2023). Um Machine-Learning-Modelle zu ermöglichen, die Varianten auf Gen-Ebene mit dem Phänotyp verbinden, wurde eine funktionale Repräsentationen von Genen entwickelt und bewertet (Brechtmann et al., 2023). Schließlich wurde die Custom-iGEx pipeline entwickelt, die eine integrative Berechnungsmethode, die genetische Daten, Genexpressionsvorhersagen und Informationen aus biomolekularen Pathways integriert, um das persönliche polygene Krankheitsrisiko zu interpretieren und die genetische Grundlage der klinischen Heterogenität zu beschreiben (Trastulla et al., 2024).

Für WP5 organisierten wir einen internationalen zweitägigen Vor-Ort-Workshop "New Horizons in Computational Regulatory Genomics" mit drei Hauptrednern und insgesamt 40 Teilnehmern (weitere Informationen hier: <https://kipoi.org/summit/>). Wir haben diesen Workshop mit einem monatlichen Webinar ergänzt, das während der vierjährigen Laufzeit des Zuschusses stattfand (<https://kipoi.org/seminar/>, 50-200 Teilnehmer pro Seminar).

4. Wissenschaftlicher und technischer Stand

In den vergangenen zehn Jahren wurden von der biomedizinischen Forschungsgemeinschaft umfangreiche Datensätze im Bereich der regulatorischen Genomik aufgebaut. Diese Datensätze haben tiefe Einsichten in die Grundlagen der Genomregulation und der Funktion häufiger und seltener genetischen Varianten geliefert. Diese Einsichten stellen heute die Basis für personalisierte Genominterpretation dar. Um diese Herausforderung zu begegnen, hat die Bioinformatik-Forschungsgemeinschaft Methoden entwickelt, die große multi-omic Datensätze und Hochdurchsatz-Perturbationsassays nutzt. Zusammen mit der stetig wachsenden Verfügbarkeit von Rechenkapazität haben diese Fortschritte zur Entwicklung zahlreicher

neuartiger Ansätzen geführt, um die funktionale Rolle von genetischen Varianten in nicht Genkodierenden Bereichen mittels computergestützter und experimenteller Methoden zu bestimmen.

Die Expertise der Gruppe von **Julien Gagneur** besteht in statistischer Modellierung von Transkriptomdaten und transkriptionellen Prozessen, insbesondere von RNA-Verarbeitung, Epigenom und Systemgenetik. **Dr. Oliver Stegle** arbeitet an der Schnittstelle von maschinellem Lernen, statistischem Rechnen und den Lebenswissenschaften. Er hat bahnbrechende Computermethoden für Molekulargenetik, Einzelzellgenomik und für die Integration von heterogenen molekularen Datensätzen und klinischen Kovarianzen entwickelt. **Dr. Michael Ziller** verfügt über umfangreiche Erfahrung auf dem Gebiet der funktionalen Charakterisierung nicht-kodierender genetischer Elemente sowie der Reprogrammierung somatischer Zellen zu pluripotenten Stammzellen und der in vitro-Differenzierung entlang der neuralen Linie. In diesem Zusammenhang hat der Antragsteller funktionale Genomik-Strategien verwendet, um z.B. Schlüsseltranskriptionsfaktoren zu identifizieren, die essentiell bei der Etablierung von verschiedenen neuronalen Vorläuferzellidentitäten sind. Ein Ausgangspunkt für die Ziele dieses Projekts ist der Kipoi-Model-Zoo (<https://kipoi.org>), eine Gemeinschaftsplattform für regulatorische Genomik, die Gagneur und Stegle in Zusammenarbeit mit Anshul Kundaje (Stanford University) entwickelt haben. Es bietet einen einheitlichen Rahmen für die Archivierung, Freigabe, den Zugriff, die Nutzung und den Aufbau von Modellen, die von der Community entwickelt wurden.

5. Zusammenarbeit mit anderen Stellen.

Weitere Zusammenarbeiten erfolgten mit Heribert Schunkert (Deutsches Zentrum für Herz-Kreislauf-Forschung DZHK), Bertram Müller-Myhsok (Max-Planck Institut für Psychiatrie), Henry Völzke (Universität Greifswald), Thomas Schulze (LMU München).

II. Eingehende Darstellung

1. Ergebnisse

AP 1. Infrastruktur. Dieses AP wurde von der Arbeitsgruppe Stegle durchgeführt

AP2. Modulare Modelle basierend auf natürlichen Variationen

Dieses AP wurde von der Arbeitsgruppe Gagneur durchgeführt.

AP3. Methoden zur Modellierung von Hochdurchsatz-Reporterassays.

Im Rahmen dieses WPs hat die Gruppe Ziller schrittweise verschiedene neue deepNN zur Vorhersage der Effekte genetischer Varianten entwickelt und sukzessive integriert. Mithilfe dieser integrierten Netzwerke war es letztendlich möglich, die Genexpression von allen Genen im menschlichen Genom für einzelne Individuen basierend auf der individuellen Genomsequenz vorherzusagen (**Abb. 2**).

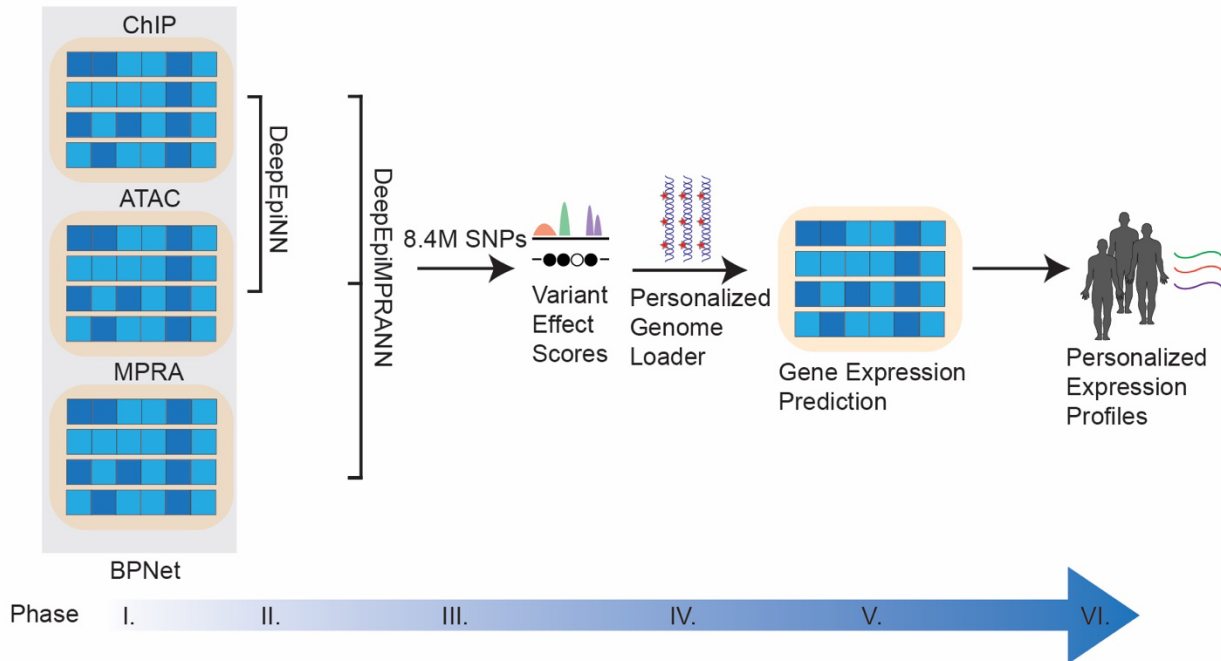


Abbildung 2 | Überblick der in AP3 trainierten deepNN Varianten und deren Integration.

In der ersten Phase wurde das zur sequenzbasierten Vorhersage von Transkriptionsfaktor-Bindung im Genom entwickelte deep learning Framework BPnet (Avsec et al. 2021) adaptiert, um auch zelltypspezifische epigenetische Profile basierend auf ChIP-Seq und ATAC-Seq Daten vorherzusagen. Zu diesem Zweck wurde ein deepNN für ChIP-Seq Daten auf über 100 Datensätzen von 80 verschiedenen Zelltypen trainiert sowie ein ATAC-Seq spezifisches deepNN, dass auf 80 ATAC-Seq Daten trainiert wurde. Für letzteres Modell wurden sowohl reguläre wie auch Einzelzell ATAC-Seq Daten genutzt. Als letztes wurde ein drittes deepNN zur Vorhersage von Hochdurchsatz Reporterassay Aktivität basierend auf im Labor erhobenen MPRA Messungen in verschiedenen neuronalen Zelltypen etabliert (Rummel et al. 2023). Sämtliche Modelle waren in der Lage, die entsprechende Modalität (ChIP-ATAC-, MPRA-Signal) mit guter Genauigkeit (mittlerer Spearman Korrelationskoeffizient Prädiktion/Messung: 0.71, **Abb.3**).

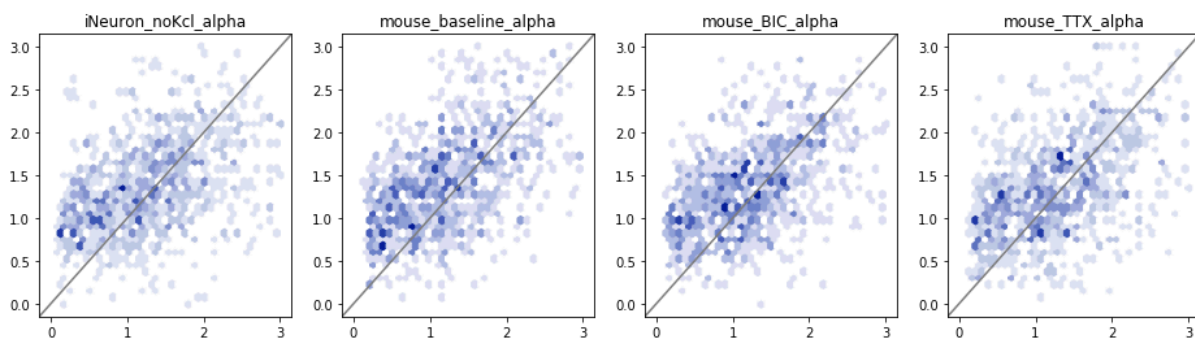


Abbildung 3 | Performance Übersicht über das MPRA-NN zur vorhersage von DNA-Sequenzaktivitäten. X-Achse zeigt die tatsächliche und die y-Achse die vorhergesagte Aktivität des regulatorischen Elements im MPRA-Assay. Der Titel gibt den jeweiligen Zelltyp in dem die Messung stattfand an.

In der zweiten und dritten Phase (Abb. 2) wurden diese drei deepNN miteinander integriert, um zunächst genetische Varianten mit einem funktionalen Effekt vorherzusagen. Diese Vorhersagen wurden dann mit existierenden, unabhängigen experimentellen Messungen des Varianten Effekts auf die Genexpression (eQTLs) validiert. Die Validierung zeigte eine hoch signifikante Anreicherung der vom kombinierten deepNN als funktional klassifizierten Varianten in den eQTLs der verschiedenen Zelltypen.

Ausgehend von diesen vielversprechenden Ergebnissen wurde in der letzten Phase (Abb. 2) das kombinierte deepEpiMPRA Modell erweitert um die personalisierte, zelltypspezifische Genexpression basierend auf dem Genotyp einzelner Individuen vorherzusagen (Abb. 4). Dazu wurde eine spezieller, hocheffizienter DataLoader entwickelt, um die für zehntausende von Individuen die personalisierte Genomsequenz effizient zur Vorhersage der Genexpression zu laden bzw. zu bearbeiten. Insgesamt werden bis zu 8.4 Millionen SNPs pro Individuum geladen und zur Vorhersage genutzt. Wie erwartet, zeigt das kombinatorische Modell eine bessere Performance als das klassische (Abb. 4). Die Ergebnisse, zusammen mit denen von AP4.3 sind Gegenstand einer Publikation in Vorbereitung.

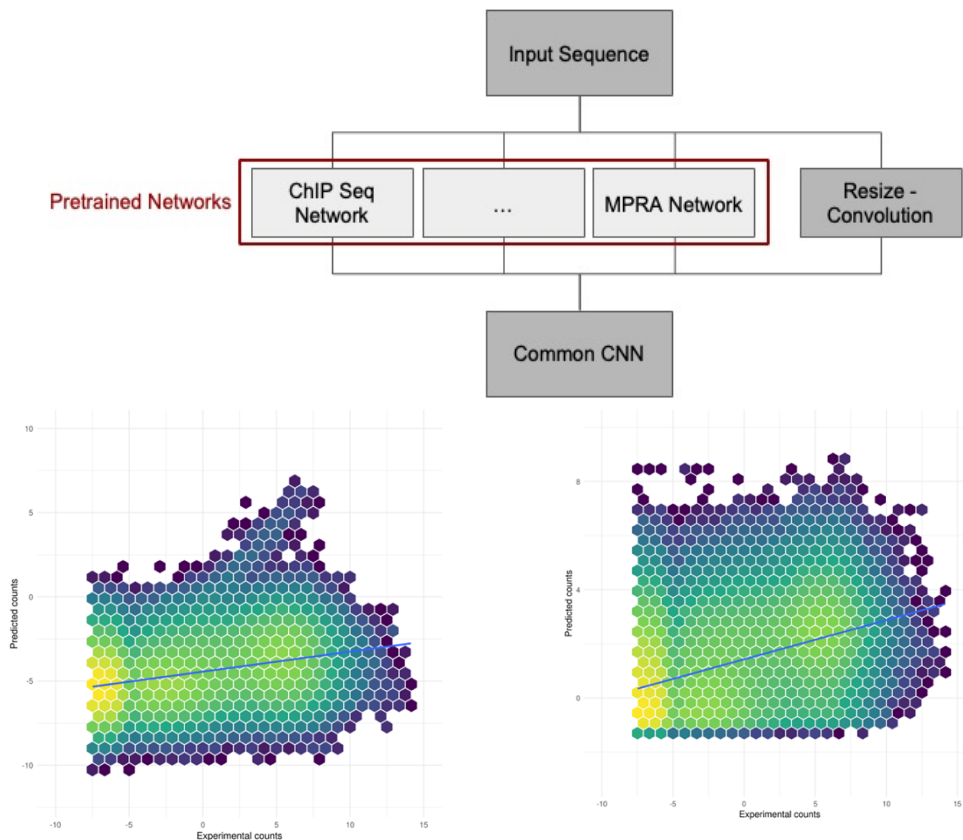


Abbildung 4 | Oben: Performance Architektur kombinatorisches deepNN, dass verschiedene vortrainierte neuronale Netzwerke im Rahmen es Kipoi Frameworks integriert um die Vorhersage von personalisierten Genexpressionsmustern zu verbessern. Unten: Ergebnisse der Vorhersage (y-Achse) und der gemessenen (x-Achse) Genexpression vorhergesagt basierend auf dem Genotyp einzelner Individuen. Die Effekte der personalisierten genetischen Varianten werden dabei entweder direkt durch ein deepNN gelernt (links, R=0.40) oder durch ein kombinatorisches deepNN vorhergesagt, welches bereits vortrainierte deepNNs basierend auf anderen genomischen Daten (ATAC-Seq, MPRA, ChIP-Seq) im Rahmen des Kipoi Frameworks integriert (rechts, R=0.48).

AP4: Vorhersage des Effekts genetischer Varianten

In einer abschließenden Studie wurde die Castom iGEx pipeline zur biologischen Interpretation des personalisierten Risikos für häufige Erkrankungen entwickelt. Castom iGEx ist eine computergestützte Methode, die genetische Daten, Genexpressionsvorhersagen und Informationen aus biomolekularen Pathways integriert, um persönliches genetisches Risiko zu interpretieren und die genetische Basis klinischer Heterogenität zu ermitteln (Trastulla et al., 2024). Die Anwendung von Castom iGEx auf Patientenkohorten mit koronarer Herzkrankheit oder Schizophrenie identifizierte verschiedene Patientenstrata oder Biotypen (**Abb. 5**). Diese Biotypen zeichnen sich durch unterschiedliche Endophänotypprofile sowie klinische Parameter wie etwa Krankheitsschwere aus. Diese verschiedenen Patientengruppen unterscheiden sich fundamental von vorherigen Stratifikationsansätzen basierend auf aggregiertem, polygenem Risiko und bieten den Vorteil, dass ihre biologische Basis direkt interpretierbar ist. Insbesondere identifiziert CASTOM-iGEx im Gegensatz zu klassischen Methoden biologisch relevante und klinisch therapierbare Patientensubgruppen, bei denen komplexe genetische Prädisposition nicht zufällig über die Individuen verteilt sind, sondern auf unterschiedliche krankheitsrelevante biologische Prozesse konzentriert werden. Diese Ergebnisse stützen das Konzept von der Existenz verschiedener Patientenbiotypen innerhalb einer Krankheitsentität, die durch teilweise unterschiedliche Pathomechanismen gekennzeichnet sind.

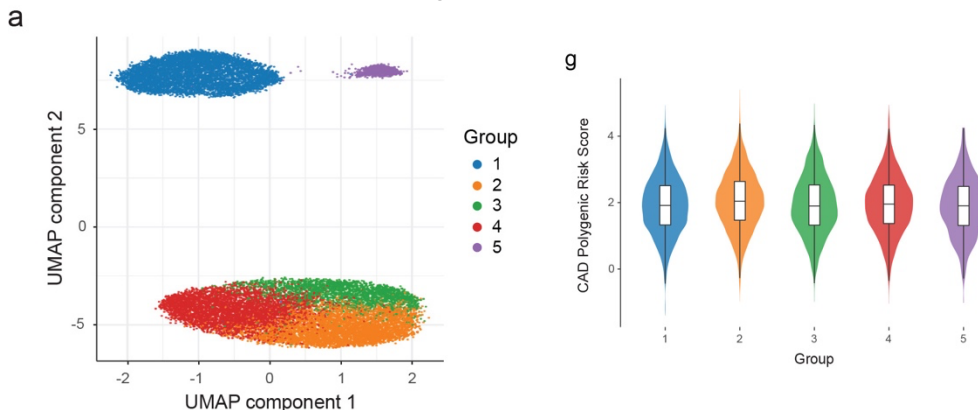


Abbildung 5 | Links: Stratifikation von 20,000 Patienten mit koronarer Herzkrankung basierend in 5 verschiedene Subgruppen auf imputierten genetischen Transkriptomprofilen in der Leber mittels CASTOM-iGEx. Rechts: Verteilung der klassischen polygenic risk score (PRS) für koronare Herzkrankung (CAD, y-Achse) über die 5 Gruppen links im Bild zeigt keinen Gruppenunterschied und unterstreicht die Neuartigkeit der identifizierten Subgruppen.

AP5 Workshop und Webinar

Bei der Einreichung des Antrags hatten wir geplant, einen Workshop zur regulatorischen Genomik zu organisieren. Leider hat die COVID-Pandemie uns daran gehindert, ihn so früh wie ursprünglich gewünscht zu organisieren. Daher haben wir ab 2020 ein monatliches Webinar eingerichtet. Das monatliche Kipoi-Seminar, das sich über die Jahren in der Wissenschaftsgemeinschaft vor großer Beliebtheit erfreut und an dem pro Vortrag jeweils zwischen 50-200 Personen teilnehmen. Die gesamte Liste der Vorträge befindet sich auf der Kipoi

Webseite: <https://kipoi.org/seminar/>. Die meisten aktuellen Vorträge werden über YouTube (https://www.youtube.com/channel/UCL_iKCHUxepOi7_Wk010--g) geteilt.

Um den Workshop organisieren zu können, hatten wir eine kostenneutrale Verlängerung um ein Jahr beantragt.

Am 25. und 26. September 2023 haben wir die Konferenz „New Horizons in Computational Regulatory Genomics“ an der Umweltforschungsstation Schneefernerhaus (<https://kipoi.org/summit/>) durchgeführt. Insgesamt 40 Wissenschaftler nahmen teil, um über Themen im Bereich maschinelles Lernen in der regulatorischen Genomik zu diskutieren. Ziel der Veranstaltung war es, neueste Forschungen zu präsentieren, Wissen auszutauschen, aktuelle Herausforderungen zu besprechen und die Zusammenarbeit zu fördern. Dieses Workshop bot auch die Gelegenheit, Ergebnisse aus unserem Konsortium vorzustellen. Die Themen umfassten:

- Transkriptionale und post-transkriptionale regulatorische Codes
- Sequenzinterpretation und -design
- Anwendung der regulatorischen Genomik auf komplexe Merkmale und Krankheiten
- Evolution genomischer Sequenzen

Darüber hinaus hatten wir eine Podiumsdiskussion über experimentelle Designs zur Entschlüsselung des regulatorischen Codes. Die Hauptreferenten waren Žiga Avsec (Google Deep Mind), Carl de Boer (University of British Columbia) und Annalisa Marsico (Helmholtz München). Die Agenda des Workshops ist auf der Website verfügbar: <https://kipoi.org/summit/>.

2. Positionen

Die Mittel wurden wie geplant eingesetzt:

Ziller

Position	Verwendung
Reise	Das Reisebudget wurde für die Teilnahme an der organisierten Konferenz ausgegeben sowie in andere Positionen umgewandelt.
Personal	Für die Entwicklung und Implementierung der statistischen und bioinformatischen Methoden wurden Doktoranden (ausgebildete Bioinformatikerin und Informatiker) angestellt. Die Mittel sind bis sind aufgebraucht worden.
Lizenzgebühren	Es wurde eine UK Biobank Lizenz erworben (siehe Verlängerungsantrag) um die Vorhersage der personalisierten Genexpression mittels deepNN für eine große Zahl von Individuen zu ermöglichen.

3. Notwendigkeit

Die im Antrag für das Teilprojekt genannten Ziele und definierten Meilensteine wurden erreicht, lediglich AP4.3 konnte im Rahmen der verbliebenen Zeit/Personalmittel nicht vollständig abgeschlossen werden. Die beantragten Mittel für Reisen, Personal und die Lizenz wurden vollständig benötigt und entsprechend eingesetzt.

4. Nutzen

Wirtschaftliche Erfolgsaussichten		Während der Laufzeit des Vorhabens	Im Anschluss an die Laufzeit
Ziel	Erläuterungen	<i>bitte zutreffende Felder ankreuzen</i>	<i>bitte zutreffende Felder ankreuzen</i>
Volkswirtschaftliche Verwertung	Die Sequenzvorhersagewerte können von Unternehmen verwendet werden.	-	x

Wissenschaftliche und/oder technische Erfolgsaussichten		Während der Laufzeit des Vorhabens	Im Anschluss an die Laufzeit
Ziel	Erläuterungen	<i>bitte zutreffende Felder ankreuzen</i>	<i>bitte zutreffende Felder ankreuzen</i>
Verbreitung der Erkenntnisse	Veröffentlichung der erzielten Ergebnisse in wissenschaftlichen Fachzeitschriften und für die breite Öffentlichkeit, Meldung bei Studienregistern, Vorträge und Poster bei Fachkongressen oder anderen Veranstaltungen	x	x

Aus-, Weiter-, Fortbildung	Vorhaben dient der Heranbildung des wissenschaftlichen Nachwuchses, durch die Erstellung von Dissertationen, die Ergebnisse finden Eingang in die Lehre	x	x
Forschungsstrukturen	Entwicklung und Instandhaltung robuster Software zur Vorhersage abweichender Transkriptionsergebnisse.	x	x

Wissenschaftliche und wirtschaftliche Anschlussfähigkeit		Im Anschluss an das Vorhaben
Ziel	Erläuterungen	<i>Zeithorizont erläutern</i>
Wissenschaftliche, technische, strukturelle und versorgungsbezogene Verwertungsmöglichkeiten	Die in diesem Projekt erzielten Fortschritte tragen wesentlich zu weiteren Forschungsprojekten und Anträgen bei: PSYCH-STRATA (https://psych-strata.eu/) und einen DFG Antrag.	PSYCH-STRATA: 10.2022-09.2027 DFG: 12.2025-11.2028

5. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Es sind von dritter Seite keine Ergebnisse bekannt geworden, die wesentlichen Einfluss auf die Verwertung der Ergebnisse nehmen.

6. Veröffentlichungen des Ergebnisses

Im Rahmen dieses Projekts wurden folgende Publikationen und Preprints von der AG Ziller veröffentlicht:

1. Lucia Trastulla, Georgii Dolgalev, Sylvain Moser, Laura T. Jiménez-Barrón, Till F. M. Andlauer, Moritz von Scheidt, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Monika Budde, Urs Heilbronner, Sergi Papiol, Alexander Teumer, Georg Homuth, Henry Völzke, Julien Gagneur, Francesco Iorio, Bertram Müller-Myhsok, Heribert Schunkert, Michael J. Ziller. Distinct genetic liability profiles define clinically relevant patient strata across common diseases. Nature Communications, 2024

Die Publikation zur deepNN basierten Vorhersage der personalisierten Genexpression befindet sich kurz vor der Einreichung.

Von Mitgliedern des Konsortiums im Rahmen des Projekts erfolgte Veröffentlichungen:

1. Pedro Tomaz da Silva, Alexander Karollus, Johannes Hingerl, Gihanna Galindez, Nils Wagner, Xavier Hernandez-Alias, Danny Incarnato, Julien Gagneur. Nucleotide dependency analysis of DNA language models reveals genomic functional elements. *bioRxiv*, 2024
2. Alexander Karollus*, Johannes Hingerl*, Dennis Gankin*, Martin Grosshauser, Kristian Klemon, Julien Gagneur. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biology*, 2024
3. Pedro Tomaz da Silva, Yujie Zhang, Evangelos Theodorakis, Laura D. Martens, Vicente A. Yépez, Vicent Pelechano, Julien Gagneur. Cellular energy regulates mRNA translation and degradation in a codon-specific manner. *Molecular Systems Biology*, 2024
4. Brian Clarke*, Eva Holtkamp*, ..., Felix Brechtmann, Florian R. Hözlwimmer, Julien Gagneur@, Oliver Stegle@. Integration of variant annotations using deep set networks boosts rare variant association genetics. *bioRxiv*, 2023
5. Florian R. Hözlwimmer, Jonas Lindner, Nils Wagner, Vicente A. Yépez, Francesco Paolo Casale, Julien Gagneur. Aberrant expression prediction across human tissues. *bioRxiv*, 2023
6. Felix Brechtmann, Thibault Bechtler, Shubhankar Londhe, Christian Mertes, and Julien Gagneur. Evaluation of input data modality choices on functional gene embeddings. *NAR Genomics and Bioinformatics*, 2023
7. Nils Wagner*, Muhammed H. Çelik*, Florian R. Hözlwimmer, Christian Mertes, Holger Prokisch, Vicente A. Yépez, Julien Gagneur. Aberrant splicing prediction across human tissues. *Nature Genetics*, 2023
8. Alexander Karollus, Thomas Mauermeier, Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 2023
9. Dennis Hecker, Fatemeh Behjati Ardakani, Alexander Karollus, Julien Gagneur, Marcel H. Schulz. The adapted Activity-By-Contact model for enhancer-gene assignment and its application to single-cell data. *Bioinformatics*, 2023
10. Jun Cheng@, Muhammed Hasan Çelik, Anshul Kundaje, Julien Gagneur@. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biology*, 2021

Literaturverzeichnis

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203. doi: 10.1038/s41592-021-01252-x

Avsec, Ž., Weilert, M., Shrikumar, A. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 53, 354–366 (2021). <https://doi.org/10.1038/s41588-021-00782-6>

- Brechtmann, F., Bechtler, T., Londhe, S., Mertes, C., and Gagneur, J. (2023). Evaluation of input data modality choices on functional gene embeddings. *NAR Genomics Bioinforma.* 5, lqad095. doi: 10.1093/nargab/lqad095
- Cheng, J., Çelik, M. H., Kundaje, A., and Gagneur, J. (2021). MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol.* 22, 94. doi: 10.1186/s13059-021-02273-7
- Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Žiga, et al. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* 20, 48. doi: 10.1186/s13059-019-1653-z
- Clarke, B., Holtkamp, E., Ozturk, H., Muck, M., Wahlberg, M., Meyer, K., et al. (2023). Integration of variant annotations using deep set networks boosts rare variant association genetics. 2023.07.12.548506. doi: 10.1101/2023.07.12.548506
- Hecker, D., Behjati Ardakani, F., Karollus, A., Gagneur, J., and Schulz, M. H. (2023). The adapted Activity-By-Contact model for enhancer–gene assignment and its application to single-cell data. *Bioinformatics* 39, btad062. doi: 10.1093/bioinformatics/btad062
- Hözlwimmer, F. R., Lindner, J., Wagner, N., Casale, F. P., Yépez, V. A., and Gagneur, J. (2023). Aberrant expression prediction across human tissues. 2023.12.04.569414. doi: 10.1101/2023.12.04.569414
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7
- Karollus, A., Hingerl, J., Gankin, D., Grosshauser, M., Klemon, K., and Gagneur, J. (2024). Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol.* 25, 83. doi: 10.1186/s13059-024-03221-x
- Karollus, A., Mauermeier, T., and Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* 24, 56. doi: 10.1186/s13059-023-02899-9
- Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., et al. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* 12, 529. doi: 10.1038/s41467-020-20573-7
- Rentsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi: 10.1093/nar/gky1016
- Silva, P. T. da, Karollus, A., Hingerl, J., Galindez, G., Wagner, N., Hernandez-Alias, X., et al. (2024). Nucleotide dependency analysis of DNA language models reveals genomic functional elements. 2024.07.27.605418. doi: 10.1101/2024.07.27.605418
- Tomaz da Silva, P., Zhang, Y., Theodorakis, E., Martens, L. D., Yépez, V. A., Pelechano, V., et al. (2024). Cellular energy regulates mRNA degradation in a codon-specific manner. *Mol. Syst. Biol.* 20, 506–520. doi: 10.1038/s44320-024-00026-9
- Trastulla, L., Dolgalev, G., Moser, S., Jiménez-Barrón, L. T., Andlauer, T. F. M., von Scheidt, M., et al. (2024). Distinct genetic liability profiles define clinically relevant patient strata across common diseases. *Nat. Commun.* 15, 5534. doi: 10.1038/s41467-024-49338-2
- Wagner, N., Çelik, M. H., Hözlwimmer, F. R., Mertes, C., Prokisch, H., Yépez, V. A., et al. (2023). Aberrant splicing prediction across human tissues. *Nat. Genet.* 55, 861–870. doi: 10.1038/s41588-023-01373-3

Münster, den 11.10.2024

Prof. Dr. Michael Ziller