GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Eberhard Karls Universität Tübingen

de.NBI – Etablierungsphase
Leistungszentrum CiBi - Zentrum für integrative Bioinformatik

Prof. Dr. Oliver Kohlbacher
Zentrum für Bioinformatik
Eberhard-Karls-Universität Tübingen
Sand 14; 72076 Tübingen

FKZ: BMBF 031 A 535A

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Fachbereich Informatik

Institut für Biomedizinische Informatik

Angewandte Bioinformatik

Prof. Dr.-Ing. Oliver Kohlbacher

Sand 14
72076 Tübingen
Germany
Telefon +49 7071 29-70457
Telefax +49 7071 29-5152
Oliver.Kohlbacher@uni-tuebingen.de
KohlbacherLab.org

Univ. Tübingen · Prof. O. Kohlbacher Sand 14 72076 Tübingen

Tübingen, den 30. Mai 2022

Schlussbericht – EKUT/OpenMS

Zu Nr. 3.2 BNBest-BMBF 98

Forschungsvorhaben:
Fkz 031A535A
de.NBI - Leistungszentrum - CIBI - The Center for Integrative Bioinformatics

Ausführende Stelle:

Prof. Dr.-Ing. Oliver Kohlbacher (Unit Coordinator)
Zentrum für Bioinformatik
Eberhard-Karls-Universität Tübingen (EKUT)
Sand 14; 72076 Tübingen
Tel. +49 7071 29 70457; Fax +49 7071 29 5152
E-Mail: oliver.kohlbacher@uni-tuebingen.de

Projektleiter:
Herr Prof. Dr.-Ing. Oliver Kohlbacher

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung, und Forschung unter dem Förderkennzeichen 031A535A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurzdarstellung

Das Ziel des "Center for Integrative Bioinformatics" (CIBI) ist die Pflege und Verstetigung der Bioinformatik-Ressourcen SeqAn, OpenMS und der Workflowumgebung KNIME, sowie die Unterstützung von Nutzern und Softwareentwicklern bei deren Verwendung. Der Standort Tübingen koordiniert die gemeinsamen CIBI Aktivitäten mit den Partnern in Berlin, Halle (Partnerprojekt), Dresden (Partnerprojekt) und Konstanz.

Der Fokus in Tübingen liegt auf der Entwicklung und Pflege der Open-Source-SoftwareOpenMS für die Analyse und Management von massenspektrometrischen Daten. Benutzer und Entwickler können innerhalb des OpenMS Frameworks Algorithmen, Tools und Workflows für die Analyse von massenspektrometrischen Daten entwickeln. Ein besonderes Augenmerk wurde bei OpenMS gelegt auf:

- die Bereitstellung grundlegender Datenstrukturen und Algorithmen für viele Anwendungsfälle aus dem Bereich der massenspektrometrischen Datenanalyse,
- die Implementierung von offenen Standards, um mit anderen Tools über gemeinsame Datenformate zu kommunizieren,
- ein gut dokumentiertes Software Development Kit (SDK),
- und eine starke Modularisierung, die es ermöglicht OpenMS Tools flexibel zu komplexen Analyseworkflows mithilfe von Workflowsystemen zu kombinieren

Um dem Umfang der Softwarelösung und den zahlreichen Anwendungsfällen Rechnung zu tragen, werden umfangreiche Schulungen, auch im ELIXIR Kontext durchgeführt.

Geleitet wird das Projekt von Prof. Oliver Kohlbacher am Lehrstuhl für Angewandte Bioinformatik an der Universität Tübingen. Die Hauptexpertise von Prof. Kohlbacher liegt in den Bereichen Algorithmen und Workflows für die Analyse von massenspektrometrischen Daten.

Des weiteren wurde durch die High Performance and Cloud Computing Gruppe, geleitet durch Dr. Jens Krüger, die de.NBI Cloud am Standort Tübingen aufgebaut und in nationale und internationale Strukturen eingebettet.

Aufgabenstellung und Voraussetzungen

Das Hauptziel des OpenMS Projekts, als Teil dieses Leistungszentrums Center for Integrative Bioinformatics (CIBI) im Deutschen Netzwerk für Bioinformatik Infrastruktur (de.NBI), ist die Wartung und Weiterentwicklung von OpenMS, um Werkzeuge für die Omics-Ebenen Metabolomik und Proteomik verfügbar zu machen. Die Entwicklung einer gemeinsamen Integrationsstrategie und technischer Grundlagen erfolgt in enger Zusammenarbeit mit den Partnerprojekten und soll die Integration diverser omics Ebenen innerhalb von Integrationsplattformen und Workflowsystemen ermöglichen. Aufgrund des Umfangs und Vielfältigkeit der eingesetzten Tools und Methoden wird in EKUT/OpenMS ein besonderer Schwerpunkt auf Training gelegt. Eine detaillierte Ausführung der erfüllten Aufgaben der einzelnen Arbeitspakete finden sich in den erzielten Ergebnissen.

Planung und Ablauf des Vorhabens

Zu Beginn des Vorhabens wurden 5 Arbeitspakete für CIBI definiert:

1. Management (Verantwortliche Stelle: EKUT)
2. Workflows (Verantwortliche Stelle: UKON)

3. OpenMS (Verantwortliche Stelle: EKUT)
4. SeqAn (Verantwortliche Stelle: FUB)
5. Training und Support (Verantwortliche Stelle: EKUT)
6. de.NBI Cloud-Standort in Tübingen (Verantwortliche Stelle: EKUT)

Das Team an der Universität Tübingen war hier vor allem in die Arbeitspakete 1,3 und 5 sowie der Cloud involviert, weshalb diese APs im Folgenden näher beschrieben werden sollen.

In der Aufstockungsphase kamen weitere, mit ELI1-6 gekennzeichnete, Elixir zugeordnete Arbeitspakete hinzu.

Arbeitspaket 1: Management

In der Aufbauphase wird EKUT die Interaktion mit der Zentralen Koordinierungsstelle von de.NBI (CCU und CAU) koordinieren. Des Weiteren ist EKUT bei der Organisation von CIBI-internen Treffen sowie von Entwickler- und Nutzertreffen der einzelnen Projekte beteiligt. Regelmäßige Treffen zielen darauf ab, einen kontinuierlichen Austausch mit den Entwickler- und Nutzergruppen zu ermöglichen, um Probleme, Anforderungen und Anpassungen an zukünftige Technologien zu identifizieren und durchzuführen. Die koordinierte Entwicklung wird die Interoperabilität der verschiedenen Werkzeuge sowie die Interoperabilität mit anderen de.NBI-Werkzeugen verbessern und Synergien zwischen den Paketen schaffen (z.B. durch gemeinsame Build-Systeme und kontinuierliche Softwareintegrationsinfrastruktur). Gemeinsame Entwickler und Anwendertreffen werden intensive Schulungen und Tutorien beinhalten, um sowohl die Entwickler als auch die Anwender zu schulen. Schließlich wird dieses Arbeitspaket Methoden und Metriken zur Bewertung der Nutzerzahlen und der Nutzerzufriedenheit von CIBI entwickeln.

Arbeitspaket 3: OpenMS

Dieses Arbeitspaket bündelt die Bemühungen um die Wartung und Integration von Werkzeugen für Proteomik, Metabolomik und andere MS-basierte Technologien in OpenMS. Unter anderem wird als Teil dieses APs Wartung für Tools angeboten und an standardisierten Datenaustauschformaten gearbeitet. Darüber hinaus wird sich das Arbeitspaket auf die Stabilisierung und Verbesserung der allgemeinen Codequalität von OpenMS konzentrieren, um dessen Wartbarkeit durch die Einführung von Techniken wie kontinuierliche Softwareintegration, kontinuierliche Bereitstellung der Software und Codereviews für die OpenMS-Entwickler und Anwender anzubieten. Durch die Bereitstellung von Dokumentation für die bestehende Codebasis wird dieses Arbeitspaket einen leichten Einstieg für externe Entwickler und Mitglieder anderer de.NBI-Standorte gewährleisten.

Arbeitspaket 5: Training und Support

Der Fokus dieses Arbeitspakets ist das Erstellen von Trainingsmaterial und das Durchführen von Trainings für verschiedene Zielgruppen: Einsteiger in die Proteomik/Metabolomik, Anwendungsbenutzer, Datenanalysten und Entwickler. Die Komplexität der Kurse reicht von kurzen Tutorials auf Konferenzen bis zu einwöchigen Vorlesungen an Universitäten. Bei der Erstellung neuer Trainingsmaterialien ist das Feedback der existierenden Community einzubeziehen und existierende Trainingsmaterialien der individuellen Projekte sollten harmonisiert und wenn möglich miteinander verknüpft werden.

Arbeitspaket 6: Cloud

Die de.NBI Cloud ist als Plattform zur Ausführung von bioinformatischen Analysen und Auswertung komplexer lebenswissenschaftlicher Datensätze konzipiert. Sie ist als föderierte Cloud angelegt, die über gemeinsame Authentifizierungs- und Authorisierungsmechanismen verfügt, die über die Life Science AAI (formals ELIXIR AAI) realisiert werden. Verbunden ist dies mit einem zentralen Management von Cloud-Projekten und den dafür notwendigen Ressourcen. Am Standort Tübingen wurde ab 2016 mit dem Aufbau der Cloud-Infrastruktur begonnen, die über die Jahre kontinuierlich erweitert wurde. Parallel wurde ab 2017 mit der Projektarbeit in SIG6 begonnen, in der äußerst erfolgreich eine gemeinsame Management- und Governance-Struktur etabliert wurde. Besonderer Fokus wurde für die de.NBI Cloud Tübingen schon sehr früh auf die bereitstellung von angemessenenen Ausführungsumgebungen für die Prozessierung sensibler Daten gelegt. Das Informationssicherheitsmanagementsystem der de.NBI Cloud Tübingen wurde im November 2021 nach ISO27001 zertifiziert. Die de.NBI Cloud Tübingen ist an EOSC-Life und HealthyCloud beteiligt, außerdem dient sie als Infrastruktur für DataPLANT und GHGA.
ELIXIR-Arbeitspakete

- ELI1 ELIXIR Workflow-Systems: Anpassung der Knotengenerierung sowie Bereitstellung der entsprechenden Tools im Galaxy ToolShed.
- ELI2 ELIXIRWorkflows: Es werden etablierte Genomics und Proteomics Workflows für Galaxy und Nextflow implementiert, getestet und global zur Verfügung gestellt.
- ELI3 und 4 ELIXIR Core-Workflows: Koordination mit der nf-core Gemeinschaft zur Integration der neuen Nextflow workflows in nf-core.
- Elixir Arbeitspaket: ELI5 ELIXIR-Documentation: Erstellen von Dokumentationen für externe Entwickler um deren SeqAn/OpenMS basierenden Tools auch in die entsprechenden Workflowsysteme zu integrieren
- Elixir Arbeitspaket: ELI6 ELIXIR-Trainings: Vorbereitung und Durchführung von User und Developer Trainings auf internationalen Workshops und Konferenzen.

Eingehende Darstellung

Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Die CIBI/Tübingen Arbeitspakete (CIBI.X.Y) ergeben sich aus dem ursprünglichen Antrag und wurden in der Aufstockungsphase um weitere Punkte erweitert. Insbesondere seit der Beteiligung von de.NBI an ELIXIR hat sich das CIBI an verschiedenen Aktivitäten auf europäischer Ebene beteiligt. Diese Arbeitspakete sind gesondert mit ELI1-6 gekennzeichnet.

In der Etablierungsphase hat EKUT die Koordination mit der CCU und CAU aufgenommen und in regelmäßige Treffen bis zum Projektende durchgeführt (CIBI.1.1). Regelmäßige CIBI workshops, interne Arbeitstreffen (CIBI.1.3) und Trainingsevents wurden etabliert (CIBI.1.2, Details in AP 5) und Entwicklungen sowie Hardware zur Erhebung projektspezifischer Metriken umgesetzt und bereitgestellt (CIBI.1.4.). Eine OpenMS Internetpräsenz (CIBI.1.6, www.OpenMS.de) wurde erstellt und Services in die de.NBI Internetseite eingepflegt (CIBI.1.5).

Während der Projektlaufzeit entwickelte sich eine enge Zusammenarbeit mit UKON/KNIME, um die OpenMS Werkzeuge in die KNIME Analytics Platform für omics Analysen zu integrieren und so gemeinsam benutzbar zu machen (CIBI.3.1). Hierzu wurden Anpassungen an OpenMS Werkzeugen nach einhergehender Codeanalyse, Qualitätskontrolle und Sichtung der Dokumentation durchgeführt.

Um sicherzustellen, dass bestehende Tools weiterhin funktionieren, wurde eine kontinuierliche Softwareintegrationsstrategie zusammen mit FUB/SeqAn entwickelt (CIBI.3.2). Des Weiteren wurde im Verlauf des Projekts die Integrationsinfrastruktur erweitert, um eine kontinuierliche Softwarebereitstellung zu ermöglichen (CIBI.3.2 z.B. in Form von nächtlichen Installern, Paketen, und Virtualisierungsimages/Containern). Allgemein wurde die Code-Dokumentation und -Qualität durch die Einbindung automatischer Checks (linting) in den Entwicklungsprozess erheblich verbessert. Insgesamt wurden sieben OpenMS Releases zur Verfügung gestellt.

Neben Erweiterung der Tutorials, Dokumentation und Trainingsmaterial für OpenMS (CIBI.5.3) und Bereitstellung dieser Ressourcen im Internet wurde die Community auch durch Veranstaltungen wie Hackathons, Trainings, User Group Meetings, Developer Meetings, Workshops und Summer Schools gepflegt (CIBI.5.4), wobei gleichsam die Sichtbarkeit des Projekts und von de.NBI erhöht wurde. Insgesamt wurden über die Projektlaufzeit 31 Veranstaltungen vom OpenMS Projekt abgehalten. Diese Kurse richteten sich vor allem an Entwickler und Benutzer des OpenMS Frameworks. Zusätzlich wurden Workshops auf zwei Summer Schools abgehalten und OpenMS war Teil der regelmäßigen CIBI User Group Meetings. Insgesamt nahmen 571 Teilnehmer an vom OpenMS Projekt organisierten Kursen teil (264 PhDs, 165 Postdocs, 10 Industrieteilnehmer, 132 Unbekannt) - 307 aus Deutschland, 265 aus dem Ausland. Durchschnittlich wurde unsere Kurse zu 92% mit “Sehr gut” oder “Exzellent” bewertet. 94% würden eure Kurse weiterempfehlen. Abgesehen von den fokussierten Trainingsevents wurden täglich Entwickler bei der Integration ihres eigenen Codes und bei der Datenprozessierung mit OpenMS unterstützt (CIBI.5.5).

Zum Aufbau der de.NBI Cloud Tübingen (CIBI.6.1) wurde in 2016 der erste Teil der Hardware ausgeschrieben. In den Folgejahren wurde ähnlich verfahren und mindestens eine EU-weite Ausschreibung durchgeführt, um die Cloud zielgerichtet nach den Bedürfnissen der Nutzenden zu erweitern. Das Gesamtinvestitionsvolumen umfasst mehr als 6 Mio. €. Die Infrastruktur am Standort Tübingen ist über zwei Standorte verteilt und bietet so eine erhöhte Ausfallsicherheit und georedundante Absicherung für die Datenhaltung. Gegenwärtig umfasst die de.NBI Cloud Tübingen 8936x CPU-Cores und 176 TB RAM. Dazu kommen 188x leistungsstarke GPUs (96 NVIDIA RTX A6000 und 92x NVIDIA Tesla V100) die sowohl die Ausführung spezifischer bioinformatischer Applikationen beschleunigen, wie auch für Anwendungsfälle unter dem Einsatz von Machine Learning genutzt werden (CIBI.6.2). An beiden Standorten werden umfangreiche Storage-Ressourcen vorgehalten, zum einen auf Quobyte basierendes multipurpose Storage wie auch eine auf Ceph basierende Langzeitspeicherlösung. Das Quobyte Storage System umfasst derzeit etwa 18.400 TB an verfügbaren HDD Speicher und etwa 645 TB an verfügbaren SSD Speicher. Der Speicherplatz der durch das Ceph System für Archivierung bzw. als Langzeitspeicher genutzt werden kann beträgt derzeit 14.760 TB (HDD) und etwa 614 TB (SSD).

Der Aufbau und Betrieb der de.NBI Cloud Tübingen (CIBI.6.1) als Teil der de.NBI Cloud Föderation wurde und wird in enger Abstimmung mit den Partnern in SIG6 und über die CCU koordiniert. Hierdurch wird die Interoperabilität mit den anderen de.NBI Standorten sichergestellt und die Kompatibilität mit anderen Cloud-Initiativen, wie EOSC, CSC Cloud, Embassy Cloud oder GAIA-X hergestellt.

Die de.NBI Cloud Tübingen unterstützt zahlreiche Initiativen und Projekte auf regionaler, nationaler und internationaler Ebene. Derzeit arbeiten an allen 86 aktiven Projekten 254 registrierte Nutzer. So ist die Cloud infrastrukturelle Basis für die NFDI-Konsortien DataPLANT und GHGA. Außerdem wird das Science Data Center BioDATEN unterstützt und mit dem baden-württembergischen HPC-DIC und Cloud Projekten eine enge Kollaboration gepflegt. Bei den europäischen Projekten EOSC-Life

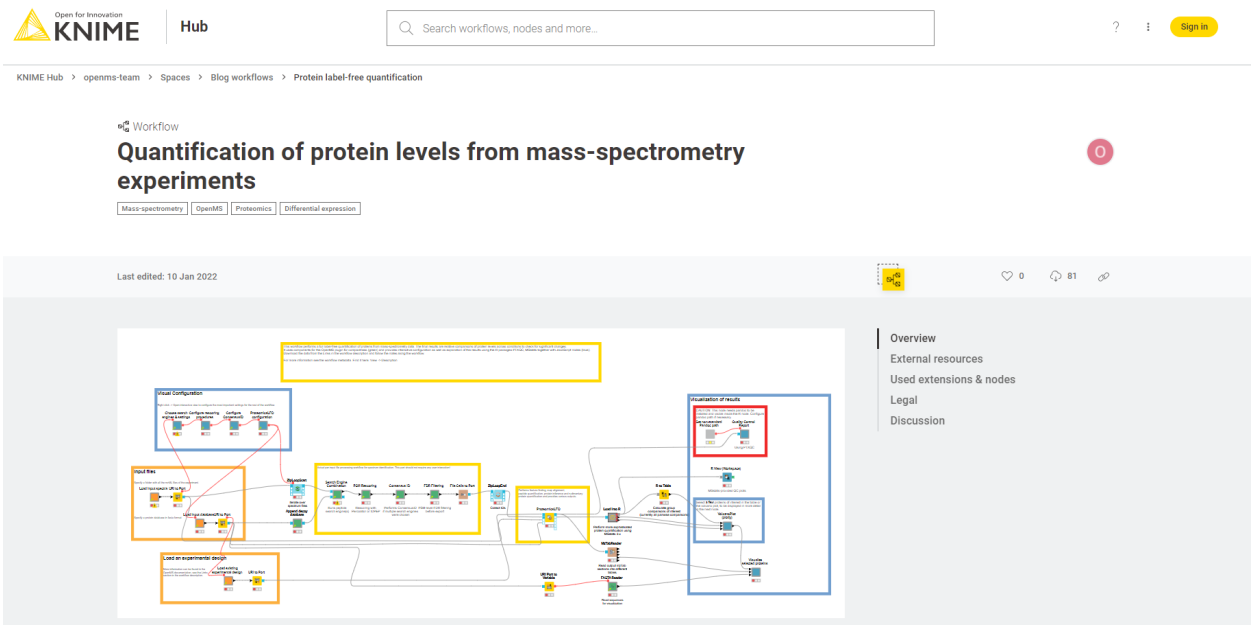
und Healthy Cloud besteht eine direkte Beteiligung, gemeinsam mit den anderen de.NBI Cloud Standorten. Darüber hinaus werden Galaxy Europe und nf-core unterstützt, was unmittelbar der Community durch Unterstützung bei bioinformatischen Workflows zugute kommt. Dazu kommen zahlreiche fachspezifische Projekte und Konsortien, die infrastrukturell und fachlich unterstützt werden. Hierzu zählen unter anderem Risikofaktoren für Parkinson in COURAGE-PD, Parkinson Epidemiologie in Lux-GIANT, Augenscreening der Nationalen Kohorte, Molekulare Epidemiologie am BNITM, Virusgenome über CoGDat, Automatisierte Image-Prozessierung mit der Radiologie UKT, Machine Learning auf Probandendaten in TREND und zahlreiche weitere. Viele dieses Projekte basieren auf der Prozessierung von sensiblen Daten für die ein besonders hohes Schutzniveau erforderlich ist. Für die de.NBI Cloud Tübingen werden mit den Forschenden individuelle Verträge zur Datenverarbeitung im Auftrag flankiert und spezifischen Standard Operating Procedures ausgearbeitet. Der zeit- und personalintensive Einsatz macht sich für alle Beteiligten bezahlt, die de.NBI Cloud Tübingen ist eine der wenigen etablierten Forschungsdateninfrastrukturen auf der sensible Daten DSGVO-konform verarbeitet werden können.

Voraussetzung hierfür ist ein resilienter Betrieb der de.NBI Cloud Tübingen und die Pflege eines effektiven Informationssicherheitsmanagementsystems (CIBI.6.3), wodurch das Vertrauen der Nutzenden auf eine solide Basis gestellt wird. Die de.NBI Cloud Tübingen, beziehungsweise ihr Informationssicherheitsmanagementsystems wurden im November 2021 nach der ISO-Norm 27001 auditiert und erfolgreich zertifiziert (CIBI.6.4). Hiermit wird nachprüfbar dokumentiert, dass die Cloud-Umgebung höchsten Standards genügt und auch für die Prozessierung von sensiblen Daten uneingeschränkt geeignet ist.

Im Rahmen des Aufstockungsantrags wurden innerhalb der bewilligten Arbeitspakete die Arbeiten zur Integration der neuen Partnerprojektstandorte, der Koordination am de.NBI-Cloud Standort, und der Koordination des erhöhten Trainingsaufkommens ausgeführt (CIBI.1.7). Im Rahmen der Aufstockung wurde eine Continuous-Deployment-Strategie am Standort Tübingen implementiert (CIBI.3.3), und mittels automatischer Checks/Tests eine Erhöhung der Codequalität, eine verbesserte Dokumentation des Quellcodes, sowie eine Erhöhung der Testabdeckung erreicht (CIBI.3.4). Des Weiteren wurden Workflows mit MetFrag und Fiji entwickelt (CIBI.3.5) und gemeinsame Trainingsevents angeboten (CIBI.5.6).

ELIXIR Aktivitäten CIBI/Tübingen

Das CIBI begrüßte den Anschluss von Deutschland an Elixir und hat Services auf europäische Ebene erweitert. Wir haben erfolgreich unsere Pläne zur einfacheren Integration von Tools in die auf europäischer Ebene weit verbreiteten Workflow Manager Nextflow und Galaxy umgesetzt. Zum einen wurde zusammen mit UFZ Leipzig an der automatischen OpenMS Galaxy Knoten Generierung gearbeitet und Hochdurchsatzworkflows in nf-core (z.B. https://nf-co.re/proteomicslfq) öffentlich gemacht. Beides Projekte, an denen bereits mehrere ELIXIR Partner beteiligt sind. Des Weiteren wurde die Entwicklung des KNIME-Hubs, einer Plattform zum Teilen von Best-Practise-Workflows und Workflowkomponenten, mit KNIME abgestimmt und Workflows auf dem KNIME Hub bzw. dem Social Workflow Repository veröffentlicht (ELI1-6).



Ein OpenMS Workflow im Social Workflow Repository

Beteiligung an sechs ELIXIR Implementierungsstudien durch CIBI/OpenMS im Rahmen von ELI 1-6:

1. “Mining the Proteome: Enabling Automated Processing and Analysis of Large-Scale Proteomics Data”, 2017
2. “Crowd-sourcing the annotation of public proteomics datasets to improve data reusability”, 2018 – 2019.
3. “Extending open proteomics data analysis pipelines in the cloud: Additional tools and focus on scalability, supporting the dramatic growth of public proteomics data”, 2018 – 2019.
4. “Comparison, benchmarking and dissemination of proteomics data analysis pipelines”, 2019
5. “Standardizing the fluxomics workflows”, 2019-2021
6. “Increasing the translational value of public proteomics datasets: Automatic metadata-driven reanalysis in cloud infrastructures”, 2021

ELIXIR-Bezug der Cloud

Als Teil der in der SIG6 mit allen Partnern ausgearbeiteten Strategie verfügt die de.NBI Cloud über eine uniforme Authentifizierungs- und Authorisierungsinfrastruktur basierend auf der Life Science AAI (formals ELIXIR AAI). Die de.NBI Cloud Tübingen ist integraler Bestandteil hiervon und hat die Entwicklungen aktiv mitgestaltet. Hierbei wird unter anderem eng mit ELIXIR Finland und ELIXIR Czech Republic zusammengearbeitet. Flankiert wird die Kollaboration zu Aktivitäten zu sensiblen Daten und ihrer Absicherung und Dokumentation, wobei auch mit dem ELIXIR EMBL-EBI zusammengearbeitet wird.

Wichtige Positionen des zahlenmäßigen Nachweises

Zu den wesentlichen Kostenfaktoren an der Universität Tübingen zählte die Beschäftigung von wissenschaftlichen Mitarbeitern sowie die Cloud. Zusätzlich gab es Ausgaben für die Beschäftigung

von Hilfswissenschaftlern, Reisekosten und Sachmittel für externe Entwicklungen. Die über die gesamte Projektlaufzeit angefallenen Kosten sind dem Verwendungsnachweis zu entnehmen.

Notwendigkeit und Angemessenheit der geleisteten Arbeit

Erklärtes Ziel des Deutschen Netzwerks für Bioinformatik-Infrastruktur ist das Bereitstellen von Infrastruktur, die es Forschenden in Deutschland ermöglicht, neue computergestützte Methoden zu entwickeln und Forschungsdaten effektiv auszuwerten. Um dies zu ermöglichen, existieren bereits verschiedenste Datenbanken, Werkzeuge, und Dienste, die jedoch unabhängig voneinander betrieben oder gefördert werden. Die an CIBI/Tübingen geleisteten Arbeiten waren über den gesamten Projektzeitraum sowohl in de.NBI als auch später in den ELIXIR-Kontext eingebettet, um in enger Entwicklungsarbeit und Trainingsvorhaben mit den jeweiligen Communities diese Ziele erreichbar zu machen. Das eingesetzte Personal und der Ressourcenaufwand waren notwendig und angemessen, um auf CIBI Ebene die oben genannten umfangreichen Arbeiten und Trainingsprogramme zu bewerkstelligen und notwendig , um auf Netzwerkebene das ambitionierte Ziel einer deutschen Bioinformatik-Infrastruktur zu ermöglichen.

Voraussichtlicher Nutzen, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Die technischen Entwicklungen während der Projektlaufzeit (Integrationen, Anpassung der APIs) stehen auch in Zukunft im Rahmen der Open-Source-Entwicklung von OpenMS zur Verfügung, werden gewartet und weiterentwickelt. Neue Entwicklungen (SeqAn+OpenMS), wie die Arbeit an einer gemeinsamen Bibliothek zur Integration von CWL (common workflow language) werden fortgesetzt und sollen einer breiten Entwicklergemeinde zur Verfügung gestellt werden. Dasselbe gilt für die Entwicklung und Integration neuer Tools in das OpenMS Framework und deren Integration in bestehende und neue Analyseworkflows. Damit wollen wir sicherstellen, dass neueste Methoden zeitnah der wissenschaftlichen Gemeinschaft verfügbar gemacht werden. Fortführung und Ausbau von Dokumentation und Training sowie wissenschaftliches Consulting werden auch in Zukunft der Gemeinschaft zur Verfügung stehen. Die während der Projektlaufzeit erzielten Ergebnisse werden allerdings nicht unmittelbar kommerziell verwertbar sein, da der wesentliche Anteil der zu erbringenden Leistungen die Betreuung von Nutzern und Softwareentwicklern bzw. die Pflege existierender Projekte sein wird. Zusätzlich sind die durch CIBI zu verstetigenden Projekte Open-Source-Projekte, was eine direkte kommerzielle Verwertung der Software nur indirekt erlaubt. Die während der Projektlaufzeit etablierten Strukturen für Support und Consulting können nach Abschluss des Projekts, vollständig oder teilweise verstetigt werden, um die Dienste auch nach Ablauf der Förderung z.B. auch in einem industriellen Kontext weiter anzubieten.

Während der Durchführung des Vorhabens dem ZE bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen

Während der Projektlaufzeit sind keine erwähnenswerten Fortschritte bekannt geworden.

Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 6.

EKUT de.NBI Personal hat mit mehr als zwanzig Veröffentlichungen einen aktiven Beitrag zur Forschung und Bekanntheit von de.NBI geleistet:

Marcu, Ana, et al. "HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy." *Journal for immunotherapy of cancer* 9.4 (2021).

Bichmann, Leon, et al. "DIAproteomics: A Multifunctional Data Analysis Pipeline for Data-Independent Acquisition Proteomics and Peptidomics." *Journal of Proteome Research* 20.7 (2021): 3758-3766.

Dai, Chengxin, et al. "A proteomics sample metadata representation for multiomics integration, and big data analysis." *bioRxiv* (2021).

Hanussek, Maximilian, et al. “Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources” *PLOS Computational Biology* https://doi.org/10.1371/journal.pcbi.1009244.

Netz, Eugen, et al. "OpenPepXL: An open-source tool for sensitive identification of cross-linked peptides in XL-MS." *Molecular & Cellular Proteomics* 19.12 (2020): 2157-2168.

Hanussek, Maximilian, et al. “BOOTABLE: Bioinformatics benchmark tool suite for applications and hardware” *Future Generation Computer Systems* (2020) 102:1016-1026.

Stützer, Alexandra, et al. "Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry." *Nature communications* 11.1 (2020): 1-12.

Starke, Robert, et al. "Tracing incorporation of heavy water into proteins for species-specific metabolic activity in complex communities." *Journal of proteomics* 222 (2020): 103791.

Alka, Oliver, et al. "OpenMS and KNIME for Mass Spectrometry Data Processing." *Processing Metabolomics and Proteomics Data with Open Software*. 2020. 201-231.

Wein, Samuel, et al. "A computational platform for high-throughput analysis of RNA sequences and modifications by mass spectrometry." *Nature communications* 11.1 (2020): 1-12.

Pfeuffer, Julianus, et al. "EPIFANY: A Method for Efficient High-Confidence Protein Inference." *Journal of proteome research* 19.3 (2020): 1060-1072.

Hulstaert, Niels, et al. "ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion." *Journal of proteome research* 19.1 (2019): 537-542.

Bichmann, Leon, et al. "MHCquant: automated and reproducible data analysis for immunopeptidomics." *Journal of proteome research* 18.11 (2019): 3876-3884.

Perez-Riverol, Y., et al. "Ten Simple Rules for Taking Advantage of Git and GitHub (vol 12, e1004947, 2016)." *PLOS COMPUTATIONAL BIOLOGY* 15.6 (2019).

Alka, Oliver, et al. "OpenMS for open source analysis of mass spectrometric data." *PeerJ Preprints* 7 (2019): e27766v1.

Hanussek, Maximilian, et al. “BOOTABLE: Bioinformatics Benchmark Tool Suite” *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Larnaca, Cyprus, (2019), pp. 157-160.

Belmann, Peter, et al. “de.NBI Cloud federation through ELIXIR AAI” *F1000Research* (2019) 8:842, doi.org/10.12688/f1000research.19013.1.

Bartusch, Felix, et al. “Reproducible Scientific Workflows for High Performance and Cloud Computing” *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Larnaca, Cyprus, (2019), pp. 161-164.

Hoffmann, Nils, et al. "mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics." *Analytical chemistry* 91.5 (2019): 3302-3310.

Perez-Riverol, Yasset, et al. "The PRIDE database and related tools and resources in 2019: improving support for quantification data." *Nucleic acids research* 47.D1 (2019): D442-D450.

Deutsch, Eric W., et al. "Expanding the use of spectral libraries in proteomics." *Journal of proteome research* 17.12 (2018): 4051-4060.

Gläßle, Benjamin, et al. “de.NBI Cloud Storage Tübingen” *in Proceedings of the bwHPC Symposium 2018*, Freiburg, 201-215, doi.org/10.15496/publikation-29062.

Kahles, André, et al. "Comprehensive analysis of alternative splicing across tumors from 8,705 patients." *Cancer cell* 34.2 (2018): 211-224.

Pfeuffer, Julianus, et al. "OpenMS–A platform for reproducible analysis of mass spectrometry data." *Journal of biotechnology* 261 (2017): 142-148.

Flett, Fiona J., et al. "Differential Enzymatic 16O/18O Labeling for the Detection of Cross-Linked Nucleic Acid–Protein Heteroconjugates." *Analytical chemistry* 89.21 (2017): 11208-11213.

da Veiga Leprevost, Felipe, et al. "BioContainers: an open-source and community-driven framework for software standardization." *Bioinformatics* 33.16 (2017): 2580-2582.

Audain, Enrique, et al. "In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics." *Journal of proteomics* 150 (2017): 170-182.

Veit, Johannes, et al. "LFQProfiler and RNPxl: open-source tools for label-free quantification and protein–RNA cross-linking integrated into proteome discoverer." *Journal of proteome research* 15.9 (2016): 3441-3448.

Röst, Hannes L., et al. "OpenMS: a flexible open-source software platform for mass spectrometry data analysis." *Nature methods* 13.9 (2016): 741-748.

Perez-Riverol, Yasset, et al. "Ten simple rules for taking advantage of Git and GitHub." *PLoS computational biology* 12.7 (2016): e1004947.

Anlage

Erfolgskontrollbericht

Beitrag zu den förderpolitischen Zielen des Förderprogramms

Das Ziel von de.NBI ist einen Verbund von Bioinformatikzentren zu etablieren, der den Zugang zu qualitativ hochwertigen Diensten sowie aktuellen Technologien ermöglichen soll und diese möglichst vielen Forschern zugänglich machen soll. CIBI bündelte hierbei das Wissen und die Erfahrung in den Bereichen der omics Technologien, um die Forschung in der Genom, Transkriptom, Metabolom und Proteomanalyse zu unterstützen. Mit der Erweiterung von de.NBI wurden zwei weitere Partner in CIBI aufgenommen, um Lücken in den Bereichen Metabolomik (Halle) und mikroskopische Bildverarbeitung (Dresden) zu schließen und Synergien im Bereich der Omicsanalyse auszubauen. CIBI/Tübingen hat als Koordinator einen wichtigen Beitrag innerhalb von CIBI übernommen und mit dem Softwareprojekt OpenMS einen wesentlichen Baustein zu Omicsanalysen, sowie zur Integration von Omicstechnologien auf Deutschland (de.NBI) und Europaebene (Elixir) beigetragen. CIBI Tübingen zeigte hohe Präsenz auf Konferenzen, hat sehr aktiv publiziert und war eines der aktivsten Knoten im Bereich Training von Wissenschaftlern im In- und Ausland. Mit der de.NBI Cloud Tübingen stellt CIBI Tübingen eine weitere zentrale Komponente der de.NBI Cloud-Infrastruktur für Forscher aus den Lebenswissenschaften zur Verfügung.

Training und Schulungen mit Beteiligung von de.NBI Personal angestellt an der Universität Tübingen

Titel des Trainingsevents	**Datum**
1st de.NBI Summer School 2021 - Analysis and integration of Mass Spectrometry based omics data in Proteomics, Metabolomics and Lipidomics	2021-09-27
CIBI User Meeting 2021	2021-09-13
Non-targeted label-free Proteomics - GCB 2021	2021-09-09
OpenMS Developer Meeting 2021	2021-04-15
OpenMS Developer Meeting 2020	2020-08-03
CIBI Data & Code Clinic - June 2020	2020-06-24
Introduction to Computational Proteomics - EuBiC 2020	2020-01-13
Label-free quantification with OpenMS - NETTAB/BBCC	2019-11-11
Protein-RNA cross-linking with RNPxl/OpenMS - SSP2019	2019-11-03
Proteomics and Metabolomics with OpenMS and pyOpenMS - GCB2019	2019-09-16
Proteomics and Metabolomics with OpenMS	2019-04-29
SeqAn & OpenMS (CIBI/de.NBI) Integration Workshop	2019-03-18

Label-free quantification with OpenMS - GCB 2018	2018-09-25
OpenMS User Meeting 2018	2018-09-19
Computational Mass Spectrometry with OpenMS - ECCB 2018	2018-09-08
Computation and statistics for mass spectrometry and proteomics	2018-04-30
OpenMS Developer Meeting 2018	2018-04-02
Analysis of Mass Spectrometry and Sequence Data with KNIME - KNIME Spring Summit	2018-03-09
de.NBI Winter School on Computational Metabolomics	2018-03-05
Developer training: Third-party tool integration and method development in OpenMS - EuBIC 2018 developer's meeting	2018-01-09
OpenMS User Meeting 2017	2017-09-25
Computational Mass Spectrometry - GCB 2017	2017-09-18
Proteomics and Metabolomics with OpenMS	2017-05-01
OpenMS & KNIME - Protemic Forum 2017	2017-04-03
OpenMS Developer Retreat 2017	2017-03-26
Introduction into the Analysis of Mass Spectrometry and Sequence Data with KNIME	2017-03-17
Label-free quantification and quality control with OpenMS	2017-01-12
9th OpenMS User Meeting	2016-09-21
Label-free quantification using OpenMS and workflows - GCB 2016	2016-09-12
OpenMS Developer Meeting 2016	2016-03-13
Von Spektren zu Ergebnissen: Effiziente Analyse von MS-Daten mit Workflows - DGMS 2016	2016-02-28
Fundamentals of Proteome Bioinformatics Revisited	2015-09-27
8th Open MS User Meeting	2015-09-16

Schulungen und Trainingsevents

Trainingsmaterial für ELIXIR Trainingskurse

https://www.denbi.de/online-training-media-library/427-lecture-computational-proteomics-and-metabolomics
https://www.denbi.de/online-training-media-library/407-introduction-to-qualitative-proteomics
https://www.denbi.de/online-training-media-library/397-introduction-to-openms

Linksammlung: Trainingsmaterial ELIXIR

Wissenschaftlich-technisches Ergebnis des Vorhabens, die erreichten Nebenergebnisse und die gesammelten wesentlichen Erfahrungen

Wichtigstes technisches Ergebnis von CIBI/EKUT sind die technischen Grundlagen für die kontinuierliche Bereitstellung gut dokumentierter und robuster Omics-Softwarelösungen im Bereich der massenspektrometrischen Proteom- und Metabolomanalyse. Die Einbindung in bestehende Workflowsysteme erlaubt nun die Hochdurchsatzanalyse und Integration mit Tools anderer Omicsbereiche. Die neuen Analyseworkflows werden durch ausführliches Trainingsmaterial für Entwickler und Anwender komplementiert. Der Betrieb des Cloudstandorts Tübingen stellt ein weiteres essenzielles Standbein für die Bereitstellung von Forschungsinfrastruktur und Rechenressourcen dar. Die gewonnenen Erfahrungen im Bereich Softwareintegration, Deployment und Testing für reproduzierbare Analyseworkflows wurden zusammen mit der Community entwickelt und erfolgreich publiziert. Weitere Publikationen mit de.NBI Personal der Universität Tübingen unterstreichen die Breite der Anwendung der entwickelten Lösungen.

Das Training von insgesamt 571 Teilnehmer aus Wissenschaft und Industrie trägt einen wichtigen Beitrag zur Weiterbildung von Wissenschaftlern bei und hat lokale Kompetenzen im Bereich Training gestärkt.

Siehe “Referenzen” für die Publikationsliste, die 24 Fachartikel umfasst.

Fortschreibung des Verwertungsplans (Schutzrechte, wirtschaftliche und wissenschaftliche Erfolgsaussichten, sowie Anschlussfähigkeit)

Arbeiten, die zu keiner Lösung geführt haben

Es wurden keine Arbeiten erwähnenswerten Umfangs ausgeführt, die nicht zielführend waren.

Einhaltung der Ausgaben- und Zeitplanung

Die Ausgaben- und Zeitplanung wurde im Wesentlichen eingehalten. Verzögerungen bei Releases wurden zeitnah aufgeholt und durch die kontinuierliche Bereitstellung von Installer/Container kompensiert.

Referenzen

Marcu, Ana, et al. "HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy." *Journal for immunotherapy of cancer* 9.4 (2021).

Bichmann, Leon, et al. "DIAproteomics: A Multifunctional Data Analysis Pipeline for Data-Independent Acquisition Proteomics and Peptidomics." *Journal of Proteome Research* 20.7 (2021): 3758-3766.

Dai, Chengxin, et al. "A proteomics sample metadata representation for multiomics integration, and big data analysis." *bioRxiv* (2021).

Hanussek, Maximilian, et al. “Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources” *PLOS Computational Biology* https://doi.org/10.1371/journal.pcbi.1009244.

Netz, Eugen, et al. "OpenPepXL: An open-source tool for sensitive identification of cross-linked peptides in XL-MS." *Molecular & Cellular Proteomics* 19.12 (2020): 2157-2168.

Hanussek, Maximilian, et al. “BOOTABLE: Bioinformatics benchmark tool suite for applications and hardware” *Future Generation Computer Systems* (2020) 102:1016-1026.

Stützer, Alexandra, et al. "Analysis of protein-DNA interactions in chromatin by UV induced cross-linking and mass spectrometry." *Nature communications* 11.1 (2020): 1-12.

Starke, Robert, et al. "Tracing incorporation of heavy water into proteins for species-specific metabolic activity in complex communities." *Journal of proteomics* 222 (2020): 103791.

Alka, Oliver, et al. "OpenMS and KNIME for Mass Spectrometry Data Processing." *Processing Metabolomics and Proteomics Data with Open Software*. 2020. 201-231.

Wein, Samuel, et al. "A computational platform for high-throughput analysis of RNA sequences and modifications by mass spectrometry." *Nature communications* 11.1 (2020): 1-12.

Pfeuffer, Julianus, et al. "EPIFANY: A Method for Efficient High-Confidence Protein Inference." *Journal of proteome research* 19.3 (2020): 1060-1072.

Hulstaert, Niels, et al. "ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion." *Journal of proteome research* 19.1 (2019): 537-542.

Bichmann, Leon, et al. "MHCquant: automated and reproducible data analysis for immunopeptidomics." *Journal of proteome research* 18.11 (2019): 3876-3884.

Perez-Riverol, Y., et al. "Ten Simple Rules for Taking Advantage of Git and GitHub (vol 12, e1004947, 2016)." *PLOS COMPUTATIONAL BIOLOGY* 15.6 (2019).

Alka, Oliver, et al. "OpenMS for open source analysis of mass spectrometric data." *PeerJ Preprints* 7 (2019): e27766v1.

Hanussek, Maximilian, et al. “BOOTABLE: Bioinformatics Benchmark Tool Suite” *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Larnaca, Cyprus, (2019), pp. 157-160.

Belmann, Peter, et al. “de.NBI Cloud federation through ELIXIR AAI” *F1000Research* (2019) 8:842, doi.org/10.12688/f1000research.19013.1.

Bartusch, Felix, et al. “Reproducible Scientific Workflows for High Performance and Cloud Computing” *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Larnaca, Cyprus, (2019), pp. 161-164.

Hoffmann, Nils, et al. "mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics." *Analytical chemistry* 91.5 (2019): 3302-3310.

Perez-Riverol, Yasset, et al. "The PRIDE database and related tools and resources in 2019: improving support for quantification data." *Nucleic acids research* 47.D1 (2019): D442-D450.

Deutsch, Eric W., et al. "Expanding the use of spectral libraries in proteomics." *Journal of proteome research* 17.12 (2018): 4051-4060.

Gläßle, Benjamin, et al. “de.NBI Cloud Storage Tübingen” *in Proceedings of the bwHPC Symposium 2018*, Freiburg, 201-215, doi.org/10.15496/publikation-29062.

Kahles, André, et al. "Comprehensive analysis of alternative splicing across tumors from 8,705 patients." *Cancer cell* 34.2 (2018): 211-224.

Pfeuffer, Julianus, et al. "OpenMS–A platform for reproducible analysis of mass spectrometry data." *Journal of biotechnology* 261 (2017): 142-148.

Flett, Fiona J., et al. "Differential Enzymatic 16O/18O Labeling for the Detection of Cross-Linked Nucleic Acid–Protein Heteroconjugates." *Analytical chemistry* 89.21 (2017): 11208-11213.

da Veiga Leprevost, Felipe, et al. "BioContainers: an open-source and community-driven framework for software standardization." *Bioinformatics* 33.16 (2017): 2580-2582.

Audain, Enrique, et al. "In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics." *Journal of proteomics* 150 (2017): 170-182.

Veit, Johannes, et al. "LFQProfiler and RNPxl: open-source tools for label-free quantification and protein–RNA cross-linking integrated into proteome discoverer." *Journal of proteome research* 15.9 (2016): 3441-3448.

Röst, Hannes L., et al. "OpenMS: a flexible open-source software platform for mass spectrometry data analysis." *Nature methods* 13.9 (2016): 741-748.

Perez-Riverol, Yasset, et al. "Ten simple rules for taking advantage of Git and GitHub." *PLoS computational biology* 12.7 (2016): e100

Berichtsblatt

1. ISBN oder ISSN geplant	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel Schlussbericht – EKUT/OpenMS	
4. Autor(en) [Name(n), Vorname(n)] Kohlbacher, Oliver Krüger, Jens Sachsenberg, Timo	5. Abschlussdatum des Vorhabens 31.12.2021
	6. Veröffentlichungsdatum
	7. Form der Publikation
8. Durchführende Institution(en) (Name, Adresse) Prof. Dr. Oliver Kohlbacher Zentrum für Bioinformatik Eberhard-Karls-Universität Tübingen (EKUT) Sand 14; 72076 Tübingen	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 031A535A
	11. Seitenzahl 16
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 1
	14. Tabellen 2
	15. Abbildungen 1
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	

18. Kurzfassung

Das Ziel des "Center for Integrative Bioinformatics" (CIBI) ist die Pflege und Verstetigung der Bioinformatik-Ressourcen SeqAn, OpenMS und der Workflowumgebung KNIME, sowie die Unterstützung von Nutzern und Softwareentwicklern bei deren Verwendung. Der Standort Tübingen koordiniert die gemeinsamen CIBI Aktivitäten mit den Partnern in Berlin, Halle (Partnerprojekt), Dresden (Partnerprojekt) und Konstanz.

Der Fokus in Tübingen lag auf der Entwicklung und Pflege der Open-Source-Software OpenMS für die Analyse und Management von massenspektrometrischen Daten.

Um dem Umfang der Softwarelösung und den zahlreichen Anwendungsfällen Rechnung zu tragen, wurden umfangreiche Schulungen, auch im ELIXIR Kontext durchgeführt.

Geleitet wird das Projekt von Prof. Dr. Oliver Kohlbacher am Lehrstuhl für Angewandte Bioinformatik an der Universität Tübingen.

Des Weiteren wurde durch die High Performance and Cloud Computing Gruppe, geleitet durch Dr. Jens Krüger, die de.NBI Cloud am Standort Tübingen aufgebaut und in nationale und internationale Strukturen eingebettet.

19. Schlagwörter

de.NBI, OpenMS, Workflow

20. Verlag	21. Preis