Towards precision medicine through integration of heterogeneous data sources

— Dissertation —

For obtaining a doctoral degree at the

Universität für Bodenkultur Wien

Submitted by

Maciej M. Kańduła

Matriculation number: 0540800

IMBT Bioinformatics Research Group

Department of Biotechnology

Advisor: **David P. Kreil**

Vienna, October 31, 2019

Acknowledgements

Firstly and most importantly, I would like to greatly thank my whole family and friends. Especially, my amazing wife Joanna, who has always been very supportive, patient and inspiring! In particular, I also thank my parents Magdalena and Wiesław for their tremendous support and enabling my studies in Vienna, and my brother Wojciech for helping out whenever needed.

I would like to thank my advisor David Kreil for giving me the opportunity to study under his ongoing mentorship, for teaching me to never settle for the minimum but always strive to achieve the maximum. Thank you for the many hours of valuable discussions and work. I have learned a lot from you and you helped me grow as a researcher, bioinformatician and as a person.

I would also like to thank Paweł Łabaj for his extensive support, both technical and scientific. Especially at the beginning of my journey with the PhD, without you it would be very hard to adapt to the then completely new environment of computational work.

For the scientific support during my stay at the Boston University in US, I would like to greatly thank Professor Eric Kolaczyk.

Also importantly, during my studies I have had the pleasure to meet and work with many wonderful people. I would like to thank all of them, especially Smriti, Norbert, and Alex for the helpful discussions, guidance and advice, and also Nancy, Nadine, and Peter.

I dedicate this work – my scientific baby of sorts, to my beloved baby son Bruno.

Abstract

Molecular life sciences advance our understanding of biology, ultimately leading to improved treatments or personalized medicine. Today, very large amounts of heterogeneous high dimensional molecular data are routinely collected from large-scale biomedical high-throughput assays. With our understanding of the different signal, bias, and noise characteristics of these data still areas of active research, the discovery of clinically relevant insight remains a bottleneck in basic and applied research, also rate-limiting the translation of research findings to the clinic. Biomedical data sets can vary substantially in size: from small patient groups in typical clinical settings, to collections of multiple cohorts spanning hundreds of cancer patients in repository databases, such as The Cancer Genome Atlas (TCGA). A joint analysis of distinct data types is increasingly explored to overcome bias and noise specific to individual studies. There is no consensus yet, however, what methods of data integration are reliably effective, with reported successes apparently often specific to a particular data set. Moreover, integrated analyses by design are more complex and can be challenging to implement and reproduce. As a result, even today, only a small percentage of published studies exploit integrated analyses.

My thesis first dissects the technical challenges of reproducibly performing complex bioinformatics analyses of high-throughput datasets, as required for successful data integration. Subsequent sections examine analyses of datasets of different molecular profile types and characteristics. I first discuss the analysis of a typical small clinical dataset and discuss the benefits and also new challenges from genome-scale assays. I subsequently consider analyses of experiments with increasing biological sample variance, from cell lines, through tissue samples from model organisms, to clinical patient samples. The increasing complexity points towards the benefits to be gained from combining different matched molecular profiles available in large cancer cohorts. I then establish and demonstrate my novel approaches to data integration on such data. In order to substantially advance methods for data integration, I also address the challenge of estimating the value and the performance of new methods in this context. Specifically, I introduce an effective number of affected patients as a balanced metric for comparing patient stratification approaches.

Through my work I show how exploration of biomolecular data can advance our understanding of biology, and human disease in particular. Recognizing that Big Data sets are often not easily understood because of their high dimensionality and heterogeneity, I show that incorporation of external knowledge and the use of similarities across data types and patients allows a meaningful dimensionality reduction that attenuates the inherent noise and facilitates the extraction of higher-level actionable patterns. In my thesis research I have designed, implemented, and validated two complementary principled methods to that end, and demonstrate their utility on a wide range of genome scale assay technologies and different cancers. One of the methods substantially improves on the state of the art of the stratification of cancer patients into clinically relevant sub-groups (ViLoN), the other method sets a new bar for patient survival prediction.

Zusammenfassung

Molekulare Biowissenschaften fördern unser Verständnis der Biologie und führen letztendlich zu verbesserten Behandlungen oder personalisierter Medizin. Heute werden sehr große Mengen an heterogenen hochdimensionalen molekularen Daten routinemäßig aus groß angelegten biomedizinischen Hochdurchsatz-Assays gesammelt. Mit unserem Verständnis der unterschiedlichen Signal-, Verzerrungs- und Rauschcharakteristika dieser Daten sind die Bereiche der aktiven Forschung noch nicht abgeschlossen, so dass die Entdeckung klinisch relevanter Erkenntnisse ein Engpass in der Grundlagen- und angewandten Forschung bleibt und die Übertragung von Forschungsergebnissen in die Klinik begrenzt wird. Biomedizinische Datensätze können sehr unterschiedlich groß sein: von kleinen Patientengruppen in typischen klinischen Umgebungen bis hin zu Sammlungen mehrerer Kohorten, die Hunderte von Krebspatienten umfassen, in Repositoriumsdatenbanken wie dem Cancer Genome Atlas (TCGA). Eine gemeinsame Analyse verschiedener Datentypen wird zunehmend untersucht, um Bias und Rauschen zu überwinden, die für einzelne Studien spezifisch sind. Es besteht jedoch noch kein Konsens darüber, welche Methoden der Datenintegration zuverlässig effektiv sind, wobei die gemeldeten Erfolge offenbar oft spezifisch für einen bestimmten Datensatz sind. Darüber hinaus sind integrierte Analysen nach Design komplexer und können eine Herausforderung bei der Umsetzung und Reproduktion darstellen. Aus diesem Grund nutzt auch heute noch nur ein kleiner Prozentsatz der veröffentlichten Studien integrierte Analysen.

In meiner Arbeit werden zunächst die technischen Herausforderungen der reproduzierbaren Durchführung komplexer bioinformatischer Analysen von Hochdurchsatz-Datensätzen, wie sie für eine erfolgreiche Datenintegration erforderlich sind, analysiert. In den folgenden Abschnitten werden Analysen von Datensätzen verschiedener molekularer Profiltypen und -eigenschaften untersucht. Zuerst bespreche ich die Analyse eines typischen kleinen klinischen Datensatzes und diskutiere die Vorteile und auch neue Herausforderungen von genomischen Assays. Anschließend betrachte ich Analysen von Experimenten mit zunehmender biologischer Probenvarianz, von Zelllinien über Gewebeproben von Modellorganismen bis hin zu klinischen Patientenproben. Die zunehmende Komplexität deutet auf die Vorteile hin, die sich aus der Kombination verschiedener angepasster Molekularprofile in großen Krebskohorten ergeben. Anschließend etabliere und demonstriere ich meine neuen Ansätze zur Datenintegration solcher Daten. Um die Methoden zur Datenintegration wesentlich voranzutreiben, stelle ich mich in diesem Zusammenhang auch der Herausforderung, den Wert und die Leistung neuer Methoden zu schätzen. Konkret stelle ich eine effektive Anzahl von betroffenen Patienten als ausgewogene Metrik für den Vergleich von Patienten-Stratifizierungsansätzen vor.

Durch meine Arbeit zeige ich, wie die Erforschung biomolekularer Daten unser Verständnis der Biologie und insbesondere der menschlichen Krankheit verbessern kann. In Anbetracht der Tatsache, dass große Datensätze aufgrund ihrer hohen Dimensionalität und Heterogenität oft nicht leicht zu verstehen sind, zeige ich, dass die Einbeziehung von externem Wissen und die

Verwendung von Ähnlichkeiten zwischen Datentypen und Patienten eine sinnvolle Dimensionalitätsreduzierung ermöglicht, die den Eigenrausch dämpft und die Extraktion von übergeordneten verwertbaren Mustern erleichtert. In meiner Doktorarbeit habe ich zwei komplementäre prinzipienbasierte Methoden zu diesem Zweck entwickelt, implementiert und validiert und ihren Nutzen an einer breiten Palette von genomischen Assay-Technologien und verschiedenen Krebsarten demonstriert. Eine der Methoden verbessert den Stand der Technik der Stratifizierung von Krebspatienten in klinisch relevante Untergruppen (ViLoN) erheblich, die andere Methode setzt einen neuen Maßstab für die Überlebensvorhersage von Patienten.

Acknowledgements i
Abstract ii
Zusammenfassung iii
List of Figures viii
List of Tables x
List of publications from my thesis research xi
1. Chapter 1: Introduction 1
1.1. Effective exploitation of growing datasets 1
1.2. Combining data for improved detection of biological signal 2
1.3. Heterogeneity of data characteristics affects integration efficacy 3
1.4. Effective benchmarking of integrative methods 3
1.5. Technical challenges for implementing large scale integrative analyses 4
2. Chapter 2: Frameworks for implementation of complex bioinformatics analyses 5
2.1. The critical role of workflow systems in Data Science 5
2.2. Workflow experiences – Use cases 6
2.2.1. Nationwide supercomputer support for sequence analysis, Sweden 6
2.2.2. Method-centric explorative research, Austria 8
2.2.3. National high performance compute infrastructure, Finland 10
2.2.4. National core genomics platform, Sweden 12
2.2.5. Government center for genotyping & sequencing research, Italy 14
2.2.6. Biomedical workflows as a service for the scientific community, Austria 17
2.2.7. Customized workflows for specialized bioinformatics applications, Bulgaria 19
2.3. Common automation strategies in bioinformatics 20
2.3.1. Basic scripting 20
2.3.2. Traditional makefiles 21
2.3.3. Scientific workflows 21
2.3.4. Key insights 22
2.4. Future workflow systems 23
3. Chapter 3: Traditional analyses (‘Small Data’) 24
3.1. An example from studying heart disease, a topical use case 24
3.1.1. Cohort characteristics 25
3.1.2. Gene regulation in isolated cardiomyocytes of male and female patients 27

3.1.3. Relationships between gene expression and ejection fraction 29
3.1.4. Sex specific differences in the response to pressure overload 32
3.2. Relevance of traditional analyses to clinical research today 34
4. Chapter 4: Genome-scale analyses 35
4.1. Big Data enable genome-scale exploration 35
4.2. Cell line assays as an example for low noise experiments 36
4.2.1. Chinese Hamster Ovary cells profiled by a novel high-performance microarray 36
4.2.1.1. Novel microarray design 38
4.2.1.2. Media supplement optimization through pathway information 39
4.2.1.3. Enhancing cell growth via detection of transcriptomic differences 42
4.2.2. Mouse embryonic stem cells for deriving thyroid progenitors via gene overexpression 44
4.2.2.1. Stem cell systems model complex disease states 45
4.2.2.2. *In vitro* might be on par with *in vivo* models in some applications 46
4.2.2.3. NKX2-1 overexpression leads to efficient thyroid derivation 47
4.3. Tissue samples from model organisms 47
4.3.1. Mouse model for finding the genetic program of differentiation to lungs 48
4.3.1.1. Finding distinct genetic program of lung through transcriptomic analysis 50
4.3.1.2. Understanding the genetic program through pathway analysis 52
4.3.1.3. Comparing *in vivo* and *in vitro* models 54
4.3.2. Sheep's secretome proteomics unravels molecular anti-inflammatory mechanisms 55
4.3.2.1. Technical feasibility of the novel sheep model 56
4.3.2.2. Ovine model supports comprehensive molecular profiling via proteomics 57
4.3.2.3. Sheep as a vehicle to understanding selected human diseases 62
4.4. Patient samples reflecting variation from non-uniform cohort structure 63
4.4.1. The role of data integration in overcoming limitations from unwanted variation 63
5. Chapter 5: Integration of heterogeneous data sources 64
5.1. A multi-layer network approach to data integration for patient stratification 64
5.1.1. Patient classification 64
5.1.1.1. Improvement through data integration 65
5.1.1.2. Network-based approaches 65
5.1.1.3. Challenges with performance assessment 66
5.1.2. A novel algorithm for the construction of patient similarity graphs 66
5.1.3. A metric for clinically relevant patient stratification 69

5.1.4. Performance in the CAMDA cancer data integration challenge 70
5.1.4.1. Robustness of patient stratification 70
5.1.4.2. Comparison to previous work 70
5.1.5. Validation on other cancer datasets 75
5.1.5.1. TCGA colon and rectal cancer 76
5.1.5.2. TCGA thyroid carcinoma 77
5.1.6. Achieved improvements in patient-stratification and recent ongoing work 78
5.2. A novel integrative framework for predicting survival time in cancer studies 78
5.2.1. The importance of vertical and horizontal data integration 79
5.2.2. Combining data from different cancers via dynamic graph databases 80
5.2.3. Integrating knowledge about molecular interactions 82
5.2.4. Linear combination of features for patient clustering 83
5.2.5. Machine learning models for survival time prediction 84
5.2.6. Building accurate survival time prediction models with data integration 86
6. Chapter 6: Conclusion: Towards precision medicine through data integration 87
6.1. Summary and outlook 90
7. Bibliography 92

List of Figures

Figure 1: Cumulative number of scientific publications (as found via PubMed) analysing TCGA datasets 3

Figure 2: Modular system for supporting in-flow enforcement of consistency constraints (prototype) 9

Figure 3: Visual representation of a user-made ChIP-seq data analysis workflow in the Chipster software 11

Figure 4: Components in the CRS4 center's automation system 15

Figure 5: Galaxy Workflow 16

Figure 6: Gene expression analysis in isolated human cardiomyocytes 28

Figure 7: Sex-specific reduction in the ejection fraction 30

Figure 8: Beta coefficients of the gene:sex interaction terms 31

Figure 9: Summary of design target properties 38

Figure 10: GeneOntology (GO)-Tree 40

Figure 11: Media supplement batches 42

Figure 12: Graphical depiction of the developed differentiation system 46

Figure 13: RNA-Seq analysis of purified Nkx2-1+ endodermal and ectodermal cell populations during mouse development 51

Figure 14: The embryonic lung specification program 53

Figure 15: Healing of adult and fetal cartilage defects 57

Figure 16: Sample correlation structure 58

Figure 17: Proteins implicated by a range of differential screening tests 59

Figure 18: Overview of the network structures exploited by the novel algorithm 68

Figure 19: Comparison of effective number of affected patients scores 72

Figure 20: Comparison of groups found by Variation of information fused Layers of Networks (ViLoN) or Similarity Network Fusion (SNF) algorithms 73

Figure 21: Bar plots showing the overlap of ViLoN-based groups with the clinical high / low risk labels 74

Figure 22: Kaplan–Meier curves showing the distinct survival profiles of groups found by the ViLoN algorithm in the Neuroblastoma dataset 75

Figure 23: Comparison of effective number of affected patients scores, in integrative analysis of colon and rectal cancer, with N-group stratification 76

Figure 24: Comparison of effective number of affected patients scores, in integrative analysis of thyroid carcinoma, with N-group stratification 77

Figure 25: Horizontal and vertical data integration 79

Figure 26: Workflow of data integration of the independent datasets, performed within our framework 82

Figure 27: The universal Tumour Integrated Clinical Feature (TICF) 83

Figure 28: Example of patients related semantically via internal and linked network 84

List of Tables

List of Tables

Table 1: Advantages and disadvantages of different categories of automation strategies for bioinformatics 20

Table 2: Baseline characteristics of the study population 26

Table 3: Echocardiographic characteristics of the study population 27

Table 4: Models with one regressor 30

Table 5: Models with interaction terms 31

Table 6: Enriched Gene Ontologies 40

Table 7: Enriched KEGG-pathways 41

Table 8: Selected relevant proteins 60

Table 9: Comparison to the best reported patient stratifications 71

Table 10: Comparisons of performance of integrative algorithms in 2-group stratification 71

Table 11: Best overall performance compared to the state-of-the-art clinical neuroblastoma grouping 73

Table 12: Aggregated results of cross-validation: with Tumour Integrated Clinical Feature (TICF); with integrative network 85

Table 13: Aggregated results of cross-validation, without the TICF; with integrative network 86

Table 14: Aggregated results of cross-validation, with the TICF; without integrative network 86

Table 15: Aggregated results of cross-validation, without the TICF; without integrative network 86

List of publications from my thesis research

This section lists the scientific papers arising from my PhD thesis research work, highlighting my specific contributions for each manuscript.

1. Spjuth *et al*. Experiences with workflows for automating data-intensive bioinformatics. Biology Direct (2015)

   I investigated various workflow systems with regard to features necessary for performing high-throughput bioinformatics analyses, with focus on my research group at BOKU. I tested these systems on multiple use-cases, including high-performance-computing on the Vienna Scientific Cluster (VSC), local servers, and groups of desktop machines.

2. Sayegh *et al*. Polarization-sensitive Optical Coherence Tomography and Conventional Retinal Imaging Strategies in Assessing Foveal Integrity in Geographic Atrophy. Investigative Ophthalmology & Visual Science (2015)

   I implemented and executed statistical analyses comparing multiple retinal imaging methods, including cross-validation with subsampling to confirm robustness towards outliers. This work provides an example of robustness analysis required in the study of ‘Small Data’ sets.

3. Shridhar *et al*. Transcriptomic changes in CHO cells after adaptation to suspension growth in protein-free medium analysed by a species-specific microarray. J Biotechnology (2017)

   I helped design a microarray assay and worked with other students in performing the actual validation experiments in the lab, including the differential expression analysis of the measurements that we generated.

4. Dame *et al*. Thyroid progenitors are robustly derived from embryonic stem cells through transient, developmental stage-specific overexpression of Nkx2-1. Stem Cell Reports (2017)

   I performed gene set enrichment analysis (GSEA) to examine molecular profiles at the KEGG pathway level, integrating measurements across genes for an effective dimensionality reduction of the studied RNA-Seq profiles. Based on my results we were able to narrow down the molecular mechanisms that facilitate thyroid derivation *via* overexpression of a specific gene.

5. Ribitsch *et al*. Fetal articular cartilage regeneration versus adult fibrocartilaginous repair: secretome proteomics unravels molecular mechanisms in an ovine model. Dis Model & Mech (2018)

   I developed, implemented, and executed comprehensive differential effect analysis of proteomics profiles in a multi-factorial experiment of wound healing in tissue samples from young and adult sheep. I implemented and combined multiple alternative analysis pipelines, from imputation of missing values to alternative normalization approaches. Notably, results are affected by algorithm choice, I identified the most simple pipeline showing good agreement with known base truths to demonstrate this new model system. I then selected a conservative threshold to focus on the strongest effects. I devised and compiled all figures summarizing the analysis that underpin the story of the manuscript and contributed the manuscript parts reporting analysis results and methods employed.

6. Ikonomou *et al*. The Genetic Program of Primordial Lung Progenitors. Nature Communications (2019) (*in press*)

   Going beyond established methods I explored a range of state-of-the-art processing and analysis pipelines for differential effect analysis of raw RNA-Seq profiles, followed by gene set enrichment analysis (GSEA). The divergent results were matched to known base truths to identify the most consistent analysis. This analysis then identified novel genes of interest for experimental *in vitro* validation. In fact, a key gene was identified and validated that elucidated the conditions triggering the genetic program of primordial lung progenitors, highlighting the role of certain signaling pathways specific to lung differentiation.

7. Mihaylov & Kańduła *et al*. A Novel Framework for Horizontal and Vertical Data Integration in Cancer Studies with Application to Survival Time Prediction Models. Biology Direct (2019) (*in press*)

   For my first collaboration independent of my PhD supervisor leading to a first-author paper, I was responsible for molecular profile analyses. Specifically, I sourced and mapped multiple complementary molecular profile types for data integration. I revised the study design to provide for an effective framework for establishing the superior performance of our novel integrative system, and refocused and rewrote the manuscript.

8. Ribitsch & Kańduła *et al*. Fetal tendon regeneration versus adult fibrous repair: proteomics deciphers molecular mechanisms. Scientific Reports (2019) (*submitted*)

   Building on and extending the analysis pipelines I had developed previously (Ribitsch *et al*. 2018) I performed comprehensive differential effect analysis of proteomics profiles in a multi-factorial experiment of wound healing in different tissue samples from young and adult sheep. Results were strongly affected by algorithm choice, and for tendon tissue, no

single pipeline could be identified as clearly superior. Only a conservative integration of multiple approaches allowed robust biological conclusions. I devised and compiled all figures summarizing the analysis that underpin the story of the manuscript and contributed the manuscript parts reporting analysis results and methods employed.

9. Gaignebet & Kańduła *et al*. Sex-specific human cardiomyocyte gene regulation in left ventricular pressure overload. Mayo Clin Proc (2019) (*submitted*)

   I performed an unbiased model search to identify molecular and clinical factors underlying improvements in heart function after surgery. This required a non-linear mapping of the target variable – the ejection fraction – and extensive robustness analysis to avoid sensitivity to outliers in a small cohort, as are common in clinical research. Even in this small cohort I was able to robustly identify a clear sex-specific effect.

10. Kańduła *et al*. ViLoN – A Multi-Layer Network Approach to Data Integration for Patient Stratification. (*in preparation*)

    I devised and established an integrative network-based algorithm that incorporates external knowledge about molecular interactions in order to find similarities between patients. I demonstrated the value of this novel approach in a typical application of clinical research, the stratification of patients for tailored treatment of cancer. Notably, other methods have so far failed to establish a widely used common method for effective data integration in this field. Scientific progress has been impeded by the lack of rigorous metrics with application relevance. For a balanced objective comparison of my novel approach and a range of independent other methods with one another and results from the scientific literature I developed and demonstrated the utility of a novel metric which accounts not only for the amount of risk change for a patient group but also the size of the patient group. This effective number of affected patients score thus forms a natural novel metric with direct relevance to the clinic. Remarkably, my novel approach, integrating multiple molecular profiles and external domain knowledge across the patients, substantially outperformed the alternative state-of-the-art algorithms, direct clinical grouping, and the most effective patient stratifications reported in the literature.

1. Chapter 1: Introduction

1.1. Effective exploitation of growing datasets

Molecular life sciences advance our understanding of biology, be that by finding genes responsible for specific healing processes[1] or biological signals responsible for cell differentiation[2]. They help us understand specific diseases, leading to finding potential treatments. This can be by discovering molecular pathways responsible for disease development, or through finding similarities between patients, to improve treatment selection. Molecular data collected in the biomedical sciences increasingly come from high-throughput experiments, and data sets are commonly of genomic scale[3,4]. Systematic analysis of gene expression, for instance, has become a key tool for biomedical research in the laboratory and clinic[5–8]. Differential expression analysis, specifically, yields hundreds to thousands of genes implicated in a specific explored condition. While individual genes often form less reliable biomarkers[9,10] exploring functionally related sets of genes can lead to discovering robustly related phenotypes[11,12]. This has also been observed for phenotypes that are of immediate clinical relevance, such as the patient response to a targeted cancer therapy[13]. Consequently, testing for the enrichment of sets of genes of known functions constitutes one of the most popular established examples of approaches for interpreting lists of implicated genes in both basic research and clinical applications[14,15].

Particularly, in the treatment of cancers, that are the leading cause of death worldwide, right after heart disorders, effective exploration that would lead to selecting an actual therapy for individual patients is challenging because of the heterogeneity of the disease. Cancer progression and treatment response can vary widely across patients. In order to predict a patient's response clinicians delineate subgroups of patients likely to react similarly. Typical criteria of classification include clinical records, such as the age at diagnosis, sex, and comorbidities. It is by now well recognized that the underlying mechanisms can vary widely across patients. Even for patients with the same specific clinical diagnosis, for instance, we find many subtypes of breast cancer[16] or adult acute myeloid leukemia (AML)[17]. Combining clinical records and histologic tests alone often cannot reliably identify the biological processes underlying a particular tumor type[18]. Increasingly, therefore, molecular markers are now incorporated to improve the prediction of therapy response and prognosis[18–21]. Common molecular markers include changes in gene activity, such as identifying characteristic gene sets or signatures[22,23], and genomic sequence variants, such as copy number changes from deletions and amplifications of certain genomic regions, like the amplification of the MYCN gene in neuroblastoma patients[24], as well as smaller changes, such as single nucleotide polymorphisms[25].

1.2. Combining data for improved detection of biological signal

The identification and interpretation of biologically relevant patterns in general, due to the not well understood bias and noise, however, still remains a bottleneck in basic and applied research, and has been rate-limiting in the translation of research findings to the clinic[26,27]. These data are highly heterogeneous, generated by various independent technologies, and highly dimensional, with thousands of genes profiled. For instance, both copy number variation and transcriptional regulation have been implicated in cancer, and are therefore systematically being profiled in key efforts like The Cancer Genome Atlas project (TCGA)[28]. In practice, however, these data are usually generated for relatively small patient cohorts. In recent years, there has been high interest in the integrated analysis of different data types from large-scale high-throughput experiments. Data integration, be that horizontal, across patients[29–31], or vertical, across assay types[31–33], should improve sensitivity to small changes in signal[29], help finding different aspects of biology from complementary information, but also help in overcoming the inevitable bias and noise inherent in individual studies[30]. Combining different assay types may capture complementary aspects of information and, investigated across patients, could help shed new light on complex relationships of interest, such as the relationship between genotype and eventual consequences for tumour progression. Complementary measurements may, for example, facilitate accurate survival time prediction.

As a consequence, new and increasingly sophisticated computational methods are being developed to better take advantage of the complex Big Data sets collected. Increasingly, methods for an integrated analysis have been applied with good results[34,35]. For example, employing both gene expression and copy number information was reported to improve clustering for subtype analysis using a probabilistic model of joint latent variables[36,37]. Notably, profiles of copy number variation, known to be a fairly noisy data source on itself[38,39], when combined with gene expression should improve predictive power and help identify prognostic biomarkers in lung cancer[40]. Evidence shows that even diagnosing a specific cancer can be extremely hard without effective integration of clinical, morphologic, immunohistochemistry, and molecular data[41]. In a further example, by introducing dimensionality reduction *via* low-rank approximation, Wu *et al*[42] reported an improved stratification of cancer patients, establishing a probabilistic model of different data types conditional on shared latent factors. Network-based algorithms, which have seen a variety of successful applications, such as the identification of dysregulated pathways[43,44] or an optimization of biotechnological processes[45] are also being applied for data integration. Using network representations Wang *et al*[46] exploit not just the complementary nature of data sources but also similarities across patients, *i.e.* combining data both vertically and horizontally, with the hope of identifying relevant patterns of the underlying biology of the disease. In summary, integration of heterogeneous data sources seems to be of great benefit in general, and of crucial importance in treatment of diseased patients, specifically.

1.3. Heterogeneity of data characteristics affects integration efficacy

It seems surprising, therefore, that data integration is not yet fully explored or well understood and still only a small percentage of studies exploit integrated analyses (Fig. 1). This may reflect the challenge of methods needing to effectively deal with the different measurement characteristics of each data type[47]. In particular, adding another data source can actually be detrimental due to the additional measurement noise introduced. The additional noise from a new data source can dominate overall results when other data sources contribute considerably more information compared to the information added by the new data source[48]. A successful robust integrated data analysis pipeline therefore needs to adapt to situations where an integration of data may be beneficial versus situations where additional noise from an additional data source just deteriorates the overall signal. Given that data integration is difficult, there is a great interest and great opportunity for improvement in developing novel integrative methods.

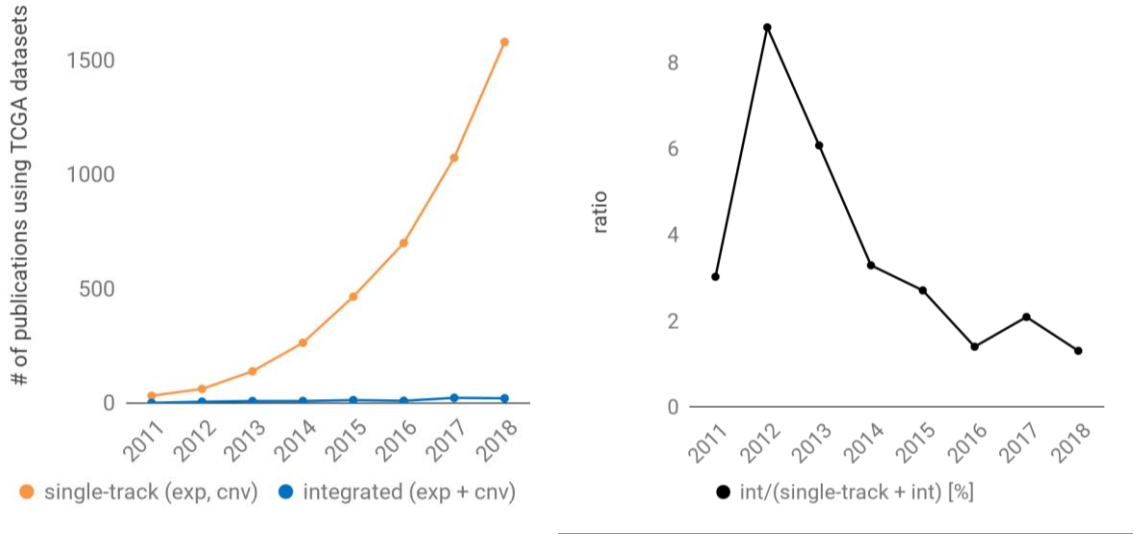

Figure 1: **Cumulative number of scientific publications (as found *via* PubMed) analysing TCGA[28] datasets**. *Left-hand* side: The number of publications per year exploring either gene expression or copy number data alone (single-track, orange line) is compared to the number of publications focusing on joint analysis of these two data sources (integrated, blue line)**.** *Right-hand* side: The decreasing ratio of published research on gene expression or copy number data analysed separately to the number of publications analysing these datasets jointly, over the years. Year 2019 is excluded as it is still ongoing.

1.4. Effective benchmarking of integrative methods

Critically, in order to objectively demonstrate that a new method is indeed advantageous it needs to be assessed through effective benchmarks in comparison with established alternatives[49–51]. Estimating the value and the performance of a new method will depend on the application scenario of interest and related factors. This means that effective benchmarking approaches need to adapt by context. For instance, in the context of predicting patient survival, commonly a group of similar patients is identified that can be associated with an average risk change. Thus a

grouping of patients is clinically relevant if one can find a sufficiently reduced or increased risk for a reasonable number of patients with significance, where significance alone is insufficient for clinical relevance[52]. Although there is no universally agreed threshold for the clinical relevance of hazard ratios for death by cancer, a risk change of 14% is typically considered to be small, changes of 47-90% can be considered moderate size effects, while risk changes of 90% or more are considered large[53]. Large effects can often be identified more easily for more specific, smaller subsets of patients. For a meaningful comparison of different predictions we therefore need to consider the number of patients as well as the size of the risk change. This is an area of active research and I will address these challenging questions in *Chapter 5*, section *A multi-layer network approach to data integration for patient stratification*.

Moreover, balanced benchmarking typically relies on both positive and negative reference sets. Especially, negative references are traditionally very hard to compile but can be very useful. For example, exploiting known negative examples has considerably improved method performance in various areas of computational biology, such as the prediction of mRNA targets of regulatory microRNAs[54] or the prediction of protein function[55], *e.g.*, for the identification of DNA-binding proteins from their sequences[56].

Because of the difficulty of determining negative samples for benchmarks, furthermore, scientists use simulated datasets, where some ground truth is known[57,58]. Non-trivial challenge for simulating data remains, however, in the necessity of capturing the biological complexity and internal structure of datasets obtained under realistic conditions[12,14,59,60].

Selecting a method amongst available alternatives has thus remained a non-trivial challenge for practitioners in biomedical research.

1.5. Technical challenges for implementing large scale integrative analyses

Also importantly, there are technical challenges related to data storage and access themselves, that need to be addressed in order to facilitate streamlined integrative analysis and validation. Data need to be accessible to an advanced algorithm. Read-only databases, like TCGA[28], that store data for download are indispensable to accelerating scientific discoveries. However, such databases relate data only by patient ids or such, and do not allow external uploading of new patient's information or simple integration with external databases. Furthermore, in such setup all data processing and analysis needs to be performed locally, including development of novel algorithms for data integration. In order to facilitate direct online interaction with the data and to make them accessible to advanced integrative algorithms, technical open issues need to be overcome. These include the problem of allowing for dynamic extension by novel input / output types, and of relating heterogeneous data sources in a meaningful and useful way[61].

Moreover, the technical complexity accompanying the development of complex integrative algorithms themselves and of performing bioinformatics analyses in general, can be reduced by workflow systems. These tools not only help to break down the intricacy of analysis, but also facilitate reproducible work − a tremendous issue in bioinformatics, and topical research area[62–66], by encapsulating reusable modules, allowing for logging of the analysis steps, versioning of software and databases, or automating reporting and documentation. Importantly, these systems facilitate performing computation on the high-performance computing infrastructure, necessary for high-throughput data analysis. Various systems have been developed[67–72], all differing in supported functionality. Even though, there exists no tool of universal appeal and applicability, there is a consensus among researchers that such systems are necessary for performing bioinformatics analyses[73].

2. Chapter 2: Frameworks for implementation of complex bioinformatics analyses

In this chapter, exploring practical use cases, I give an overview of common bioinformatics research questions, computational tasks, and supporting infrastructure and tools, with a special focus on scientific workflow systems for automated analysis[73]. All the discussed points are crucial for performing complex bioinformatics analyses, including data integration.

2.1. The critical role of workflow systems in Data Science

High-throughput technologies, such as next-generation sequencing (NGS), have revolutionized molecular biology and transformed it into a data-intensive discipline[74]. Bioinformaticians are nowadays required to interact with computational infrastructure consisting of high-performance computing (HPC) resources, large-scale storage, and a vibrant ecosystem of bioinformatics tools. It is common that analyses consist of multiple software tools applied in a sequential fashion on input data; and these analysis steps are usually executed on a server or a computer cluster given the significant data size and computation time requirements. Such a multi-step procedure is commonly referred to as a workflow. In order to efficiently carry out such analysis it can be beneficial to use Scientific Workflow Management Systems that can streamline the design and execution of workflows and pipelines in high-performance computing settings such as clusters or computing clouds[75].

There exist a number of workflow systems for use in bioinformatics. Taverna[76] pioneered integration of web services in bioinformatics; Galaxy[77–79] is a workflow system that has been used in sequence analysis and other bioinformatics applications; Kepler[80] and Chipster[81] are other examples of such systems that are used for next-generation sequencing and gene expression data analysis. All of the above mentioned systems have graphical user interfaces for constructing workflows and can run on HPC and cloud systems. However, experienced bioinformaticians

commonly work at a lower programming level and write their workflows as custom scripts in a scripting language such as Bash, Perl or Python. For this user group, a number of lightweight workflow systems have emerged to simplify scripting and parallelizing tasks on HPC resources, such as Luigi (https://github.com/spotify/luigi), Bpipe[70], Snakemake[68] and BcBio (https://github.com/chapmanb/bcbio-nextgen). General Linux tools such as Make[82,83] are also widely used due to their simplicity.

HPC resources in academia traditionally consist of compute clusters with Linux operating system and batch (queueing) systems for scheduling jobs. Recently, cloud computing has emerged as an additional technology, offering virtualized environments and the capability to run custom virtual machine images (VMI). For workflows this opens new possibilities such as packaging entire analyses or pipelines as VMIs, which has been acknowledged in bioinformatics[84]. There are also other technologies such as MapReduce[85], Hadoop[86] and Spark[87] that show great promise in bioinformatics and that might change how bioinformatics analysis can be automated.

Nevertheless, there does not seem to be a single system that satisfies the varied needs of the scientific community. To explore this further, within the COST Action BM1006: Next Generation Sequencing Data Analysis Network ("SeqAhead", http://www.seqahead.eu/), a series of hackathons and workshops brought together a number of scientists from different organizations, all involved in data-intensive bioinformatics analysis. Below, from the point of view of the participant, we summarize their current computational infrastructure and experiences with workflows, including also the challenges in automating data-intensive bioinformatics analysis recognized by specific labs. Based on these examples, I can then discuss common approaches and remaining issues, letting us propose criteria for simple and efficient systems supporting the construction and execution of complex bioinformatics in research. Specifically, I have developed an extension to a widely used workflow system (Snakemake[68]) that underpins much of my thesis work.

2.2. Workflow experiences – Use cases

2.2.1. Nationwide supercomputer support for sequence analysis, Sweden

2.2.1.1. Overview

The Bioinformatics platform at UPPMAX and Science for Life Laboratory (SciLifeLab) provide high-performance computational resources for the national NGS community in Sweden, as well as the necessary tools and competences to enable Swedish bioinformaticians to work efficiently with HPC systems[88]. Since 2010, UPPMAX has had over 500 projects and 300 users, and as of December 2014 has 3328 compute cores and almost 7 PB of storage. On UPPMAX HPC systems, users get access to installed software, reference data, and are able to carry out data-intensive bioinformatics analyses. Applications include whole genome-, de novo- and exome

sequencing, targeted resequencing, single nucleotide polymorphisms (SNPs) discovery, gene expression and methylation analysis.

2.2.1.2. Workflow experience

On our systems, most users use scripting in Bash, Perl, and Python to automate analysis. We have a security policy to not allow web servers, which has made it more difficult for us to use graphical platforms such as Galaxy. Recently, however, we have deployed a private cloud where we aim to provision images containing workflow systems like Galaxy, Chipster, and GPCR-ModSim[89], which we believe will enable us to reach a larger scientific community. We are experimenting with the workflow system Luigi on our HPC system, and CloudGene[90] on a previously established prototype Hadoop cluster in a private cloud. For automating workflow execution we use either cron jobs and an external Jenkins continuous integration instance. Besides the workflow evaluations, considerable efforts were put on the quantitative comparison of the different approaches to solve usual bioinformatic tasks in DNA and RNA-seq experiments. In recent work we provide evidence for superior scalability for the task of mapping short reads followed by calling variants on the Hadoop-with-HDFS platform compared with the existing HPC cluster infrastructure[91]. We also developed a versatile solution[92] for the feature-counting and quality assessment tasks in RNA-seq analysis, extending the acknowledged HTSeq package[93] into the e-Science domain with Hadoop and MapReduce. We are also evaluating the Spark platform for pipelining NGS data but our initial assessment did not reveal any performance gain compare to Hadoop due to the non-iterative nature of our problems. Spark has however in our opinion a more intuitive and appealing programming environment.

2.2.1.3. Future challenges

It is important for UPPMAX as a national provider of HPC resources for NGS analysis to strive for efficient resource usage. With many biologists having little experience of automating bioinformatics analyses, it is important for us to provide workflow systems, examples, support, and training in order to maximize resource utilization and improve efficiency of analyses. We are noting that future pipelines will have problems running on our current HPC systems due to intensive use of shared file systems, and we will continue to evaluate and develop a future e-infrastructure where Hadoop and Spark are interesting options. There is however a challenge for traditional HPC centers like UPPMAX to adopt cloud computing and Hadoop clusters as they contrast a lot to current best practices and experiences of system administrators. The buildup of competence in these directions will be an important task.

2.2.2. Method-centric explorative research, Austria

2.2.2.1. Overview

The Bioinformatics Research Group at the Institute for Molecular Biotechnology at BOKU University, Vienna is a method-centric research group at the interface of computational analysis and large-scale experimental assays. Recent work includes (i) an assessment of accuracy, reproducibility, and information content of gene transcript expression profiling platforms, including RNA-Seq, microarrays, and qPCR[94]; (ii) a method benchmark in the comparison of normalization efficiency across multi-site RNA-Seq laboratories[95]); (iii) signal level models of hybridization based assays for high-density microarrays[96]. These analyses require high computational power largely provided by HPC facilities like the Vienna Scientific Cluster (VSC), with the VSC-2 consisting of 1,314 nodes with 16 cores each and 32 GB RAM, and the VSC-3 consisting of 2,020 nodes with 16 cores each and 64 GB RAM. Large memory tasks are run on individual fat nodes with 256 GB–16 TB RAM.

2.2.2.2. Workflow experience

In many instances, we simply use Make[83] to run custom pipelines for both cluster and local jobs. It is a standalone tool with no setup / installation needed on standard. In our experience, if a workflow system is less lightweight than Make and small scripts (Perl, Bash, *etc.*), people will not use it when they need to 'get something done' even though many people know that in the long-term this is not efficient. Systems like Galaxy and Taverna provide a useful platform for the automation of routine data analysis steps as commonly found in industrial or facility settings, but are less useful for performing explorative and flexible analyses. In explorative work, one would like to run workflows for different configurations, and compare results. It would be helpful if there was support for tagging, managing 'alternative' workflow runs, and outputs. Moreover, what most systems lack is support for the enforcement of quality control (*e.g.*, inputs/outputs), and support for cycle control (revisions of workflows, input data, tools).

We have initially tested several systems, including, Bpipe[70], Moa (http://mfiers.github.io/Moa/index.html) and Ruffus[97], and are now focussing on exploring Snakemake[68] due to, among other features, its make-like workflow definition, simple integration with Python, Bash code portability, ease of porting workflows to a cluster, intuitive parallelization and continuous curation/development. We are currently working on extending Snakemake with a lightweight modular system for development cycle control and policy-based specification of rules and requirements that supports an in-flow enforcement of consistency constraints. We have developed a proof-of-concept prototype of the mechanism and automated the code generation of rules (Fig. 2) (unpublished data).

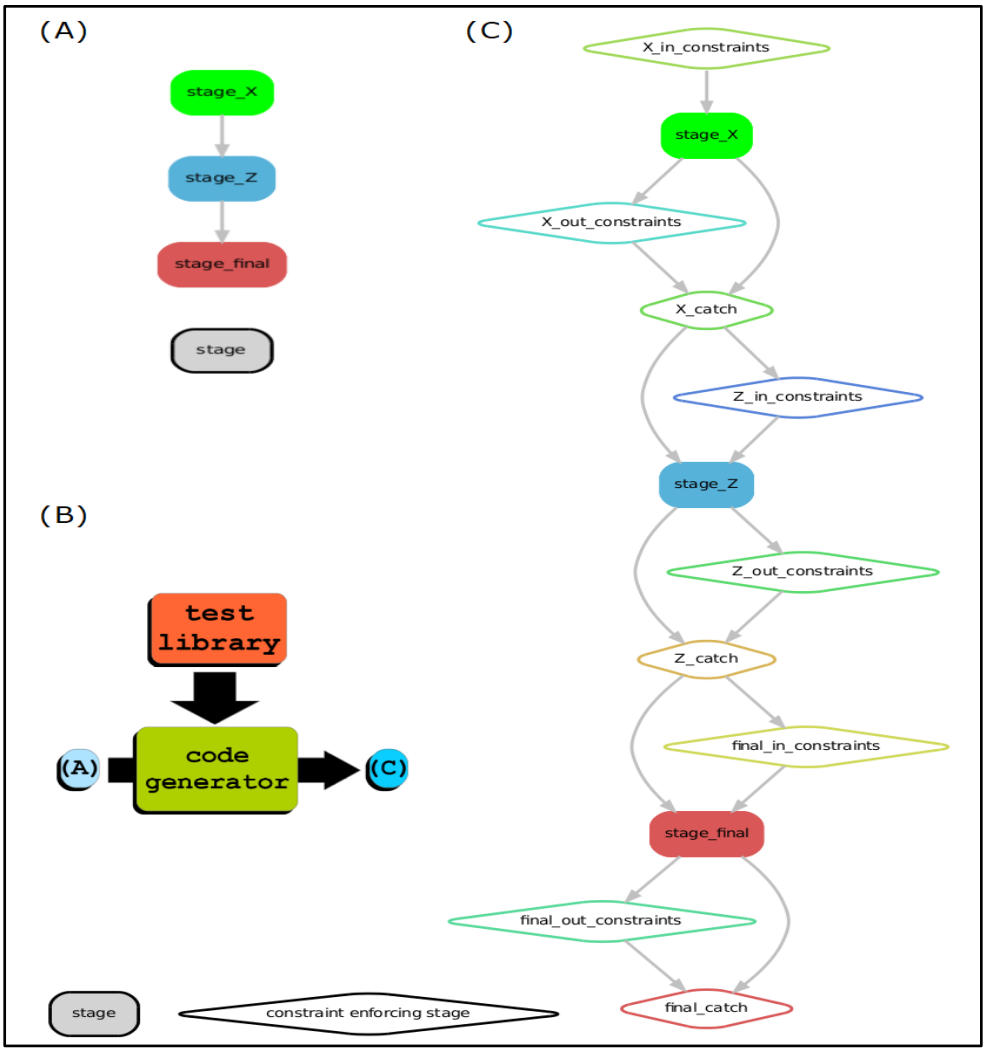

Figure 2: **Modular system for supporting in-flow enforcement of consistency constraints (prototype)**.
(A) Original directed acyclic graph (DAG) of an analysis workflow using Snakemake.
(B) Activation of constraints by combining test modules and workflow code. Steps enforcing specific constraints on input and output of a specific step of the original DAG (see A) are automatically added by our code generator. They are recognised as standard Snakemake workflow pseudo-code by the workflow system.
(C) DAG of workflow with automatically integrated constraint checking steps. All the steps (original and automatically added ones) are executed in the same fashion by the Snakemake workflow system. Here consistency constraints are enforced by our system. Specifically, these steps ensure that each analysis step is provided with non-erroneous input data, and that the data generated by this step is also correct and can be processed further.

Specifically, we use workflow systems to partly preprocess cancer-related data, like tumour/normal samples from the TCGA consortium[98] and to fully automate some steps of this data analysis. Furthermore, we try to incorporate workflows in the process of designing microarrays for Drosophila melanogaster experiments (unpublished) or to automate RNA-seq analysis in stem cell research area[2,99].

2.2.2.3. Future challenges

While Snakemake seems to be a promising tool, on its own it does not provide an allround workflow system solution, requiring external mechanisms to support critical features like revision control and management of multiple workflow instances run with varying parameter sets. We are now working to integrate Snakemake with external tools and our modular code generation system for in-flow enforcement of consistency constraints.

2.2.3. National high performance compute infrastructure, Finland

2.2.3.1. Overview

CSC - IT Center for Science is a government-owned computing centre in Finland that provides IT support and resources for academia, research institutes and companies. CSC provides capacity through a traditional batch oriented HPC environment, but also with a cloud platform. Major HPC environments are Cray XC40 supercomputer with 40,608 cores and HP XL230a cluster with 12,960 cores. The OpenStack based infrastructure-as-a-service (IaaS) cloud runs on the HPC cluster hardware.

As a national bioinformatics facility CSC has a large number of users, the majority of which have bio/medical background and no experience in programming. We strive to enable users to work independently by providing training and user friendly interfaces. An example of the latter is the Chipster software, developed at CSC, that provides a graphical user interface to a large suite of analysis tools[81].

2.2.3.2. Workflow experience

Chipster enables users to create and share bioinformatics workflows. It tracks what the user does and allows him/her to save any series of analysis steps. These workflows can be exported, shared, and applied to a different dataset. Everything is tracked, including parameter settings and reference data. The result files are also automatically annotated with this information. An example of a Chipster workflow is shown in Fig. 3.

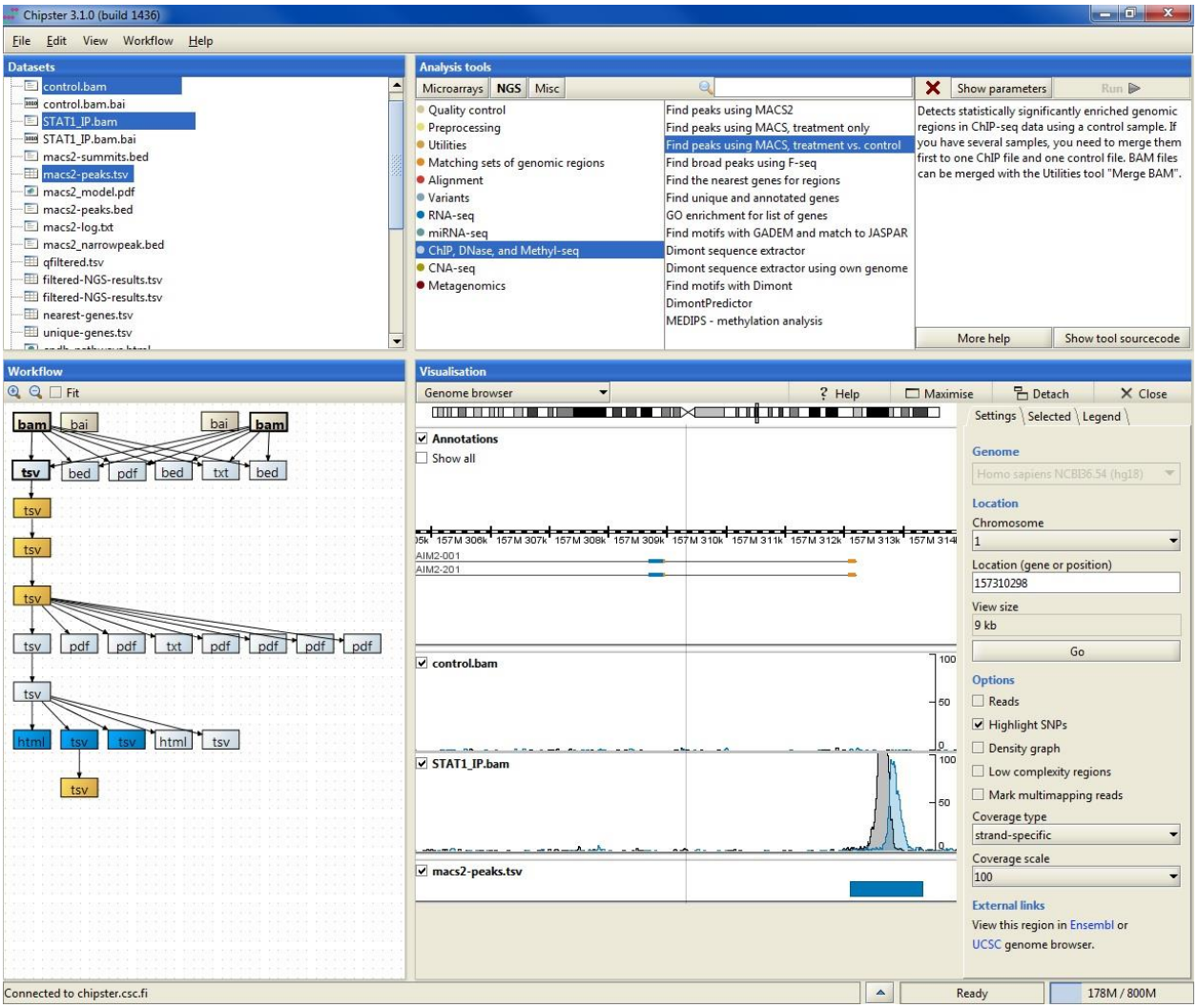

Figure 3: **Visual representation of a user-made ChIP-seq data analysis workflow in the Chipster software**. After detecting STAT1 binding regions in the genome, the user has filtered the resulting peaks for q-value, length and peak height. S/he has then looked for common sequence motifs in the peaks and matched them against a transcription factor binding site database. S/he has also retrieved the closest genes to the peaks and performed pathway enrichment analysis for them. Finally, s/he has checked if the enriched pathways contain the STAT signaling pathway. All these downstream analysis steps can be saved as an automatic workflow, which can be shared and executed on another dataset. In addition to analysing data and building workflows, Chipster allows users to visualize data interactively. As an example, genome browser visualization is shown (bottom right panel)The number of publications per year exploring either gene expression or copy number data alone (single-track, orange line) is compared to the number of publications focusing on joint analysis of these two data sources (integrated, blue line)**.** *Right-hand* side: The decreasing ratio of published research on gene expression or copy number data analysed separately to the number of publications analysing these datasets jointly, over the years.

One major challenge is where to stop when recording analysis execution. We include parameters, inputs and such, but also source code for the tools. However, maintaining full reproducibility over years is impossible because the underlying tools and databases change. Our philosophy has been to maintain reproducibility to the level that is needed for workflows to be a practical tool for users. For provenance and long term archival we store enough metadata on the workflow and,

most importantly, all data with their relationships. That might not be enough for one-click rerun of the pipeline several years later, but it is still enough for manual reproduction of the analysis.

Chipster users represent a wide range of research fields, ranging from medicine to agriculture and biotechnology. Therefore also the workflow functionality has to be flexible enough to cater for very different types of analysis. The typical tasks include analysis of RNA-seq data (QC, preprocessing, alignment, quantitation, differential expression analysis, filtering and pathway analysis), ChIP-seq data (QC, preprocessing, alignment, peak calling, filtering, motif discovery and pathway analysis) and exome/genome-seq data (QC, preprocessing, alignment,variant calling and filtering).

2.2.3.3. Future challenges

Potential future development at CSC is to provide a more technically oriented workflow engine on top of our cloud IaaS offering. We are looking into software packages that are used and developed in the cloud and big data communities to base our own development efforts on. Workflow system would be presented with platform-as-a-service (PaaS) model. Technically capable users could program workflows that are run in the IaaS cloud, but they would not need to care about the IaaS aspects such as node provisioning and user management.

Important requirement for future workflow systems is the ability to distribute data processing workload with frameworks such as Hadoop and Spark. To this end, we have participated in development of tools that allow bioinformatics data to be efficiently processed in Hadoop: Hadoop-BAM and SeqPig[100,101]. This work is continued by integrating Hadoop and Spark into our IaaS environment and providing easy to use interfaces for data intensive computing.

2.2.4. National core genomics platform, Sweden

2.2.4.1. Overview

The Stockholm genomics core platform of the Swedish National Genomics Infrastructure (NGI) crunched over 45TBp (terabasepairs) in 2014. The current NGS instrumentation located in Stockholm includes 11 Illumina HiSeq 2500 sequencers, 3 MiSeq systems, and 3 HiSeq X sequencers, and with the coming addition of more HiSeq X instruments, the amount of data produced and processed at NGI is expected to increase dramatically in the year ahead.

2.2.4.2. Workflow experience

NGI in Stockholm uses bcbio-nextgen (https://github.com/chapmanb/bcbio-nextgen) and some customizations for assembling and running the analysis pipelines. For us, having support from a pipeline framework already established in other institutions has been a big plus. In our experience, home-grown bioinformatics pipeline frameworks not published or released early

enough in the development process fail to gain wide adoption and momentum. As bioinformatics pipelines are inherently complex, we think it is better to share this complexity with the open source community and generalize as early as possible. Unfortunately we have not been able to keep up with fast developments upstream and periodically deploy validated instances of the pipeline.

We think that this shows the growing disconnect between traditional HPC architectures in academia and other sectors in industry:

1. *Non-community maintained software*. Such as using the ancient, hard to maintain and update "module system" (http://modules.sf.net) versus a more sustainable option such as the HomeBrew science (https://github.com/chapmanb/homebrew-cbl) system.

2. *Non-existent stable usage of cloud computing architectures*. This could enable continuous integration and delivery. Having containerized execution units coupled with good software management would increase robustness and provenance tracking on pipelines. That is, globally trackable software releases as opposed to the home-grown local module system that we now use.

3. *Lack of career paths for Research Software Engineers* (RSE) personnel (http://www.rse.ac.uk/who.html) that could explore new avenues and maintain points 1 and 2. In other words, lack of a "research computing" unit able to keep up and be up to date with new ways of computing.

For instance, our current HPC system does not now (and is not predicted to anytime soon) support newer deployment strategies such as continuous deployment of lightweight Docker containers (https://github.com/chapmanb/bcbio-nextgen-vm). As a result, we are actively exploring workflow frameworks and methodologies that can survive the age of HPC systems. We are investigating Piper (https://github.com/johandahlberg/piper), Snakemake, and Luigi, which seem to be more adaptable with regard to deployment strategies.

On the one hand, many pipelines incorporate a basic test suite to ensure that all moving parts work as expected. On the other hand, few of those include a benchmarking suite that can validate several bioinformatic tools and compare their performance and biological relevance. Bcbio-nextgen has put some good care in validating that the underlying biology remains sound across software versions by following up with the "Genome in a Bottle Consortium", a gold standard for validation.

Having a continuously deployed and benchmarked pipeline allows researchers and RSEs to validate every single change in the source code, like industry does with continuous software delivery and deployment models. In this way, both source code and biology can be validated and errors spotted earlier[102]. Likewise, performance of variant callers can be continuously, closely assessed and improved quantitatively in different versions of the whole system.

2.2.4.3. Best practice pipeline

For a few years, bcbionextgen has been processing samples for the so called "best practice" pipeline at SciLifeLab. The typical outputs of the pipeline include:

- Quality assessment via FastQC.
- Contamination screening via fastqscreen.
- Alignment against preconfigured reference genomes and its indexes (mainly hg19).
- Variant analysis using the GATK toolkit and FreeBayes.
- Functional annotation of variants using SNPeff.
- Several RNAseq packages such as cufflinks and DEXSeq.

In practice, although the outputs are appreciated by service customers, there are many sample and project-specific details that have to be taken into consideration. This limits our ability to generalize the data that can be most useful to our scientists, but we found that at least the quality assessment and some alignment and coverage metrics are immediately useful to researchers.

2.2.4.4. Future challenges

Modernizing the current computing environment to more modern ways to isolate and reproduce workflows (Docker) while collaboratively managing scientific software (Homebrew Science, http://planemo.readthedocs.org/en/latest/) are big challenges that hinder reproducibility and portability. Currently, we think that systems like Piper and others are too tightly coupled with specific environments, compromising its generalization and portability.

2.2.5. Government center for genotyping & sequencing research, Italy

2.2.5.1. Overview

Center for advanced studies, research and development in Sardinia (CRS4) is a government research center with a focus on applied computing and biology. It hosts a high-throughput genotyping and sequencing facility that is directly connected to the center's computational resources (3000 cores, 4.5 PB storage). With three Illumina HiSeq2000 and two older Illumina Genome Analyzer IIx, it is the largest NGS platform in Italy. CRS4 directly participates in large-scale population-wide genetic studies – for instance, pertaining to autoimmune diseases and longevity[103,104] – and provides sequencing services for external collaborators and clients. All the data produced by the sequencing laboratory undergoes some degree of processing in the computing center, spanning from quality control and packaging to reference mapping and variant

calling. Over the past five years, the facility has processed more than 2000 whole-genome resequencing samples, 800 RNA-Seq samples and 200 exome sequencing samples.

2.2.5.2. Workflow experience

At CRS4 we have worked to automate the standard preliminary analysis of sequencing data to achieve high sample throughput and consistency. The processing system is summarized by the schematic diagram in figure Fig. 4. Our automation strategy is split in two layers. At the lower layer we are using the Galaxy platform to implement workflows for specific operations on data – e.g., demultiplexing (Fig. 5), alignment and variant calling. At a higher level, a custom daemon launches and monitors the execution of these workflows according to its configuration. When a workflow completes its operations, the daemon registers the resulting datasets in our OMERO.biobank[105] traceability framework, which allows us to keep track of which input datasets and sequence of operations were applied to produce the results (represented by serializing the galaxy history). The process effectively results in a dataset graph rooted at the original raw data.

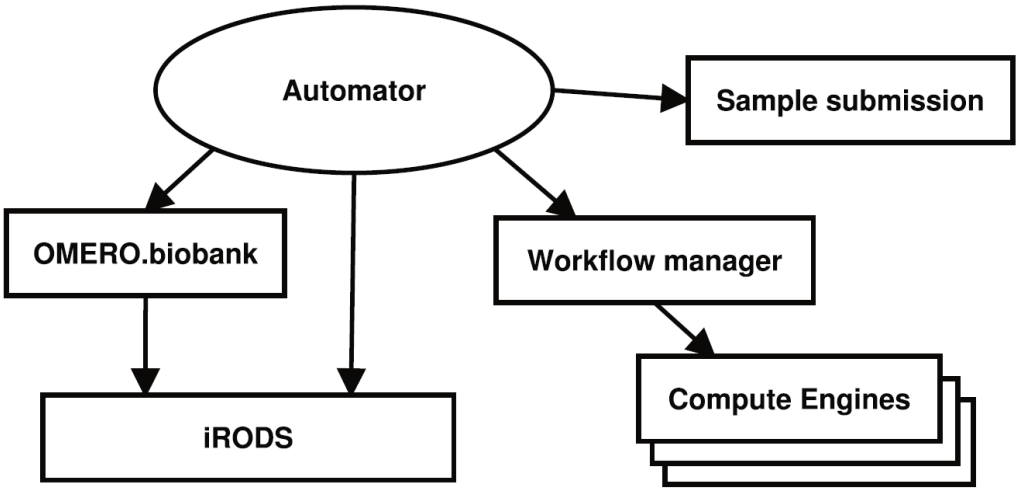

Figure 4: **Components in the CRS4 center's automation system**. The system has been created by linking together freely available components with some specialized software built in-house. In addition to running preliminary processing, it records operations within OMERO.biobank, thus ensuring reproducibility.

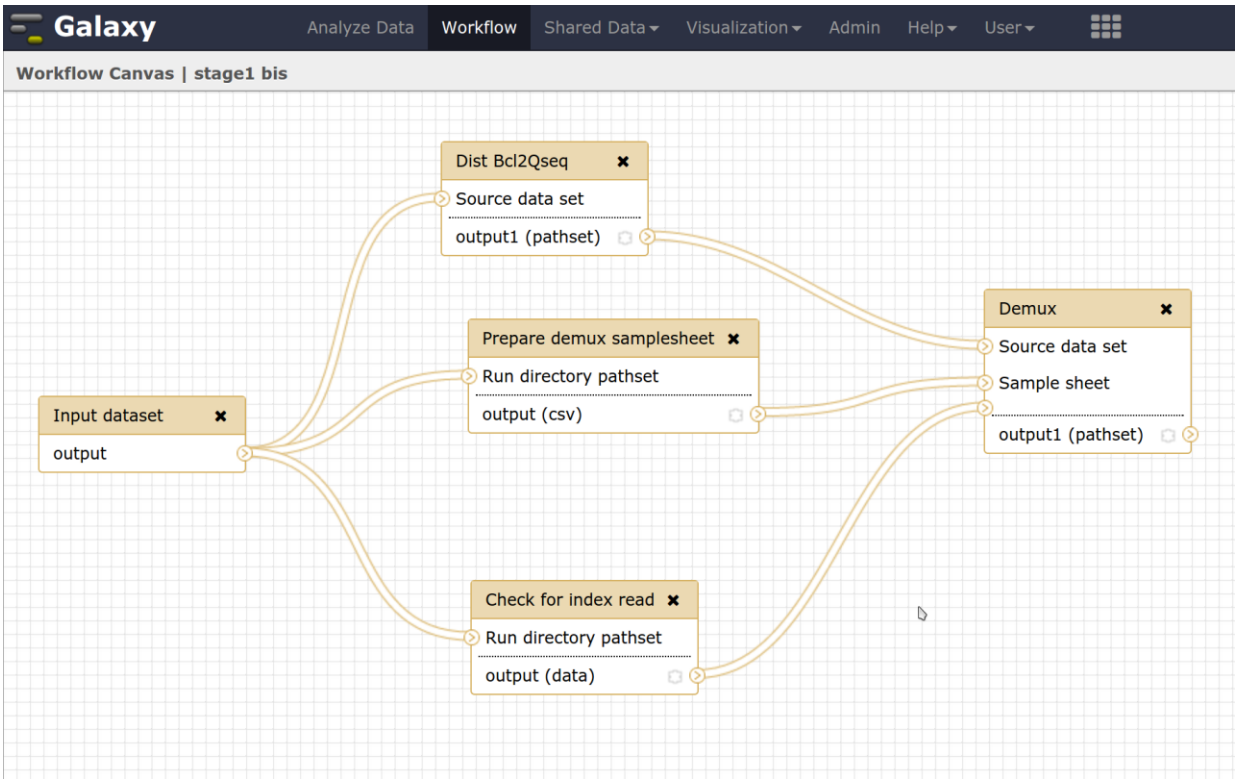

Figure 5: **Galaxy Workflow**. Example of a workflow used at CRS4 to generate demultiplexed fastq files starting from an Illumina run directory. The BCL to qseq conversion and the demultiplexing operations are performed on a Hadoop cluster using the Seal toolkit.

The automation daemon also connects multiple workflow operations in sequence, when necessary; for instance, after running the demultiplexing workflow it is configured to run a sample-specific workflow to process each sample dataset. The daemon implements an event-driven model, where events are emitted in the system when something specific happens (*e.g.*, flowcell ready, workflow finished, *etc.*) and the system is programmed to react to each event type with a specific action. The action may perform some housekeeping tasks, such as moving files to a specific location, or execute some other workflow.

To help our operation sustain a high throughput level – and to leverage CRS4's computing cluster – we implemented some of the more time-consuming and data-intensive processing steps on the Hadoop platform[106], and proceeded to integrate these tools with Galaxy[107] to compose them with other conventional tools in our bioinformatics workflows.

In summary, our operation uses Galaxy to define complex operations (workflows) given its familiarity to biologists and bioinformaticians and its REST API, which allows is to supplement it with our own custom automation daemon. On the other hand, we have turned to Hadoop-based tools to improve our computational scalability. Finally, to ensure reproducibility we trace all our automated operations with OMERO.biobank. The entire operation is described in more detail in Cuccuru *et al*[105].

2.2.5.3. Future challenges

Future challenges vary in complexity and ambition. At a lower, perhaps simpler, level lies the need to have full reproducibility of these data analysis procedures. To a degree at CRS4 we have achieved this goal by tracing all automated operations with the combination of Galaxy and the OMERO.biobank. However, the system only works with operations that are run and monitored by our automation daemon; therefore, it cannot trace interactive, user-driven operations. In addition, our current solution introduces some complexity in managing changes in workflows and tools versions. For these issues we currently rely on Galaxy, but its functionality in these terms is limited so alternative solutions will need to be devised or integrated.

A more ambitious challenge lies in the need to be able to efficiently deal with a steady stream of updates to model data (such as genomic references), bioinformatics tools and analysis procedures. To stay current, all acquired datasets need to be kept in line with the state-of-the-art. This is a very laborious, complex and computationally intensive task which, however, could be automated with the proper support for operating on both data and workflows computationally as first class citizens. With such functionality one could, for instance, update an alignment workflow to use the latest genome reference and automatically find all datasets that had been generated with the previous version and recompute them.

2.2.6. Biomedical workflows as a service for the scientific community, Austria

2.2.6.1. Overview

The Medical University of Innsbruck (MUI) is one of the leading centres of medicine in Austria. Within the MUI, the Division of Genetic Epidemiology is an internationally recognized expert on lipid-associated disorders, holds cooperations with several epidemiological studies and is involved in several genome-wide association studies (GWAS) and imputation projects. An intensive cooperation with the research group Databases and Information Systems (DBIS) at the University of Innsbruck exists, developing data-intensive bioinformatics software solutions such as Cloudgene[90], HaploGrep[108] or the mtDNA-Server[109]. Lately, the developed workflow system Cloudgene has been utilized as the underlying architecture for the Michigan Imputation Server, developed in cooperation with the Department of Biostatistics, University of Michigan. For in-house data analysis, a cloud approach based on a shared-nothing cluster architecture is used for data processing.

2.2.6.2. Workflow experience

Our institute is especially experienced in providing biomedical workflows as a service to everyone (SaaS). For example, the Michigan Imputation Server

(https://imputationserver.sph.umich.edu) provides an efficient, user-friendly and free service to impute large-scale population studies using the 1000 Genomes Panel (Phase 1 and 3) or the new HRC Panel. Furthermore, the mtDNA-Server (http://mtdna-server.uibk.ac.at) enables a highly parallelized way to detect heteroplasmies and contamination within mtDNA samples.

For these time-intensive manipulation and analysis of huge datasets, we mainly focus on the application of Hadoop (hadoop.apache.org). Therefore we developed Cloudgene, a framework for the execution and tracking of Hadoop MapReduce workflows. This graphical workflow system allows domain experts to run implemented MapReduce workflows directly from their web browsers. Cloudgene is able to combine existing MapReduce programs written in Java, approaches based on the high-level language Apache Pig, command line tools and R-based scripts to a sophisticated workflow. All used parameters and input/output data are tracked ensuring reproducibility and transparency. Final reports are created using R and RMarkdown. Within Cloudgene, workflow steps are defined in a YAML manifest file, the underlying workflow definition language (WDL) supports conditions and loops. Based on this workflow definition, Cloudgene creates user interfaces to submit MapReduce jobs graphically. Since Cloudgene supports the execution of command-line programs and bash scripts, it can also be seen as a generic workflow system. Furthermore, the architecture behind Cloudgene was developed in a way that it is compatible with existing cloud managers such as CloudMan[110]. Thus, the same workflow can be executed on a local infrastructure or on private and public clouds without any adaptations[111]. This enables us to develop prototypes of new bioinformatics workflows fast and to provide them as services to other scientists.

2.2.6.3. Future challenges

Reproducibility of data and software is from our perspective one of the most challenging tasks in the near future. Many publications are presenting software solutions, which are often hard to integrate into a local workflow ore impossible to use due to specific requirements on software packages. We think that cloud-based SaaS approaches applying state-of-the-art pipelines could improve the quality of current data analyses. Of course, Apache Hadoop is not applicable to all kinds of problems, but its scalable and open-source nature could result in a boost within Bioinformatics. One goal of Cloudgene is therefore to improve it to an even more generic big data platform by supporting the complete Hadoop YARN architecture. This opens the door to build and execute workflows based on different computational models such as Apache Spark (http://spark.apache.org) or Apache Giraph (http://giraph.apache.org).

2.2.7. Customized workflows for specialized bioinformatics applications, Bulgaria

2.2.7.1. Overview

At the Faculty of Mathematics and Informatics and Joint Genomic Center, both part of Sofia University, we do research projects that require customized workflows. This has been necessary for tasks ranging from biodiversity estimation in metagenomics, to alternative transcript detection in wheat, maize, sorghum and arabidopsis genomes, as well as SNP discovery in wheat. For this reason, graphical or web-based workflow software designed for easy creation and maintenance of workflows does not suffice for our requirements. We started with shell scripts and then moved to standard Makefiles and as an alternative, our own Python-based bioinformatics workflow system.

2.2.7.2. Workflow experience

In our experience, more modern workflow systems do not always offer significant advantages. In our bioinformatics projects, the use of Makefiles alone is not enough. During tasks such as NGS assembly, alignment or variant calling, some custom data processing which cannot be implemented in shell pipelines is usually implemented in AWK, Bash or Python scripts. AWK allows compact presentation of simple data processing and is enough in surprisingly many cases. Biopython library has also proved to be very convenient for more complex handling of bioinformatics data files.

While easy to use and construct, Makefiles are often not flexible enough - their support for parallel jobs cannot take multi-threaded or multi-process jobs into account, and they do not provide any usable means to describe a recursive flow, such as progressive application of multiple alignment for large datasets. Some of the shortcomings can be overcome by using sub-Makefiles, however we thought it would be useful to develop a YAML-based workflow description system inspired by Makefiles. We apply them for the more convoluted problems, but we are hoping to make it generally applicable to simple problems as well[112].

2.2.7.3. Future challenges

Our aim is to build a workflow tool that is as simple as Makefiles, yet one that can make use of more complex functionality. The major challenge is making the system feature-complete and as expressive as Makefiles without sacrificing the simplicity that is inherent in the alternatives. Work is also ongoing to optimize the workflow schema and extension syntax to take maximum advantage of the YAML format. We are looking for expanding our expertise in workflow systems and improving our own tool.

2.3. Common automation strategies in bioinformatics

The approaches for automating bioinformatics analysis by the organizations at the SeqAhead hackathons and workshops roughly fall into the following categories: scripting (usually in languages such as Bash, Perl, or Python), Makefiles or similar (Make, Snakemake, CMake, *etc*.), and other workflow systems (such as Luigi, Galaxy, Taverna, and BcBio). We summarize the main advantages and disadvantages from our point of view in Tab. 1. One observation is that sequencing workflows are used in two different ways. There are core workflows that are used for preliminary processing, are standardized and rarely change, and are the ones that stand to benefit the most from sophisticated automation. Then there are research workflows that a bioinformatician creates to run ad hoc analysis, explore the data and try to extract information. These are not standardized and indeed the steps and parameters are chosen and modified often as the understanding of the data and the problem changes. We note that several of the organizations are not satisfied with the currently available tools and have resided to developing in-house tools to support their use cases. We also observe that workflow tools developed and used in other domains, such as astronomy[113] in many cases are not widely used in bioinformatics, which may partly be due to a lack of communication between scientists of different fields, yet also reflect domain specific needs.

Table 1: **Advantages and disadvantages of different categories of automation strategies for bioinformatics**

	Advantages	Disadvantages
Scripting	• Simple to construct	• Hard to hand over, manual tools integration and difficult HPC interaction
Makefile	• Simple to construct once you are familiar with the programming languages and the bioinformatics command-line tools involved	• Multithreaded programs and remote execution not handled well
	• Describes data flow and takes care of dependency resolution, parallel execution and caching results from previous runs	• Lack of recursion support
	• Uses code fragments in familiar scripting languages for processing of data	• Requires programming or shell experience
		• Can't be automatically parsed and visualized
Scientific Workflow Systems	• More powerful features, easier to maintain and share	• Requires more effort

2.3.1. Basic scripting

Shell scripts are compact and tailored for running commands in a specific order. Linux has simple yet powerful commands for text processing, and most of bioinformatics tools are available as Linux commands. There are some significant disadvantages to this simple approach to automation. One is that it can be tricky to ensure reproducibility of analyses; or rather, the onus is completely on the individuals who are using the script to document in some way the datasets that have been produced. Moreover, desirable advanced features such as resilience to hardware problems, the ability to re-use intermediate datasets, integration with HPC cluster resources, etc. must all be written from scratch. It can be argued that by the time such features have been

integrated into the script one has effectively written a new workflow system, and thus might have been better off adopting one from the start.

However, scripting also has advantages and hence many adopters. The most important convenience is likely its simplicity and flexibility, meaning that one can very quickly achieve some degree of process automation that works, though it may not be optimal or efficient. Another important advantage is that most bioinformaticians already have scripting experience and are familiar with some scripting languages. By automating through scripts that knowledge can easily be recycled. In the authors' experience scripting is not sufficient to provide a fault-tolerant automation for production use.

2.3.2. Traditional makefiles

The standard Linux solution for automating compilation and other tasks that require a dependency graph are Makefiles[82]. These can serve as a simple yet effective tool to describe bioinformatics workflows, and are applicable to a wide variety of tasks. They describe dependencies between files and commands, and commands can be executed in parallel. Subsequent runs of the workflow use as many of the computation files from previous runs as possible, which serves as a basic form of caching. Drawbacks of makefiles are their inability to create multiple output files and lack of support for multiple wildcards in I/O names. Moreover, their design is not suitable for large datasets, long running operations, and execution on heterogeneous failure-prone distributed resources. To address these issues both general purpose Make implementations, like SCons (http://www.scons.org), PGMake[114] or GXP make[115] were developed, and bioinformatics-dedicated systems, like Makeflow[116] and Snakemake[68]. These tools try to move beyond Makefiles while retaining the simplicity of GNU Make [83].

As Makefiles grow they tend to become very complex. In the authors' experience, Makefiles are good for simpler use cases, but have shortcomings when it comes to more complex workflows with multiple steps and branches.

2.3.3. Scientific workflows

Scientific workflow systems provide an environment to interconnect components and in most cases allow for execution on distributed resources. Authors' experiences regarding their utility vary. While all acknowledge the power and importance of scientific workflow systems to enable reproducible data analysis and simplified integration with HPC systems, in practice it turns out that many projects have started using workflow tools and frameworks but later switched back to custom scripting and Makefiles (or similar) since they discovered limitations of the systems, especially with the pressure from PIs to deliver results faster.

An important remaining challenge is the standardization of data flow in workflow systems. There have been several attempts to address this issue, where some are based on describing common

data types via a dedicated XML schema[117] or introducing ontology-based methods for managing data types[118]. No particular approach, however, has yet emerged that could substantially impact the field or find widespread acceptance in the bioinformatics community. With no central authority to dictate standards for interoperability, the community can only develop standards through collaborative efforts like the EU COST action SeqAhead.

2.3.4. Key insights

- Automation on shared HPC clusters is difficult, and workflow tools can aid in achieving it.
- True analysis reproducibility is typically hard, sometimes impossible to achieve. This has two reasons: i) large scale analysis very often relies on external databases that commonly are not versioned, or even if they are versioned only 'milestone' versions are available, ii) scientific software management is on one hand inefficient in HPC clusters while on the other hand usage of the Web Services might be risky due to instability and lacking versioning. Community efforts for standardized software packaging and versioning are also lacking.
- The available log processing and provenance systems are not good enough. These would provide better reproducibility, monitoring and analytics.
- Bioinformatics analyses are currently to large extent file-based and there is no standardized way of passing data between applications in the workflow. This would require a transparent conversion of data formats with the resulting technical as well as semantic challenges[119]. In addition there is a necessity to check the consistency of the produced data. Although these tasks do not form a 'research' part of the workflow they can still constitute the majority of the workload in a typical analysis[120].
- Biological validation of workflows is typically missing. In other words, integration with a reference biological dataset, such as genome in a bottle, and accompanying test suite that validates biology across each source code change is, unfortunately, not a common practice today.
- Makefiles are a quick way to get the work done in a seemingly efficient manner, but they can become inadequate when more advanced features are required. It can also be difficult to understand for inexperienced users. Further, there exist more efficient tools than Makefiles with similar level of complexity (*e.g.*, Snakemake, Bpipe).
- Scripting is common for analysis development, but we see a move to Workflow tools for data production that has more strict requirements.

- Workflow systems on HPC resources have advantageous performance over cloud computing resources, but software installations is simplified on cloud systems and they are also more suitable for interactive use.

With regard to seamlessly and reproducibly performing bioinformatics analyses researchers seem to have similar problems. Especially in case of data-intensive bioinformatics analyses, where an additional layer of computational complexity is added. To overcome the discussed issues workflow tools are becoming increasingly more powerful, user-friendly, and hence more frequently used and appreciated for automation and creation of research pipelines.

2.4. Future workflow systems

The harder it is for a scientist to use a system compared to an *ad-hoc* hack, script, or perhaps a suboptimal stand-alone tool, the lower the widespread acceptance of a workflow system is in the wider bioinformatics and computational biology community. In general, we therefore recommend further development of light-weight and layered systems, where at least the basic functionality is easily accessed. More specifically:

- Maintain as much reproducibility as possible without sacrificing usability and simplicity of design and execution.
- Keep things simple, light-weight, easy to install and integrate with Bash and scripting languages.
- Workflows should be easily executed, with little or no change in local and distributed environments (HPC and cloud).
- Encourage attempts for further data flow standardization and data versioning as well as standardized software management.
- Put more effort into (biological) testing, validation, continuous delivery and deployment of the software. In other words, spend more effort on quality assurance.

Bioinformatics analysis are currently to a large extent file-based, and as long as this will be the case, workflow tools will continue to be important for bioinformatics automation. Even though exciting new data analytics frameworks such as Hadoop and Spark provide alternatives, with high up-front costs and the so far low uptake in the bioinformatics community we do not see a shift in paradigm within the nearest years.

The analyses presented in the following chapters, even if this is not explicitly stated, were facilitated *via* workflow systems, if not the whole analysis pipeline then at least parts of the bioinformatics work.

3. Chapter 3: Traditional analyses (‘Small Data’)

In clinical studies it is common to focus on a selected small group of patients for which clinical information and various clinical outcomes are directly collected by the physicians and scientists involved in the treatment[121,122]. In addition, expression of genes of interest might be measured with low-throughput technologies, like quantitative polymerase chain reaction (qPCR)[122]. This results in a low dimensional dataset, consisting of a small number of samples and only a few variables describing them. Such datasets are typical of traditional molecular biology and are the backbone of most of our current knowledge. It is known that findings presented in such studies may have low statistical power[123]. Only stringent robustness analysis, as I perform in the following subsection, allows identification of strong signals[122].

3.1. An example from studying heart disease, a topical use case

With heart and circulatory diseases being the leading causes of death worldwide, I first focus on analysis of a related dataset. In particular, here, we focus on aortic stenosis (AS) patients. The pressure overload (PO) that develops in aortic stenosis or hypertension leads to left ventricular (LV) hypertrophy (LVH). The initial compensatory mechanism is expected to reduce wall stress and to maintain systolic function of the heart. In later stages, the myocardium undergoes pathologic molecular, cellular and tissue changes. This maladaptive LV remodeling includes reactivation of the fetal gene program, induction of fibrosis, dilation, ultimately contributing to the development of heart failure (HF).

The development of PO-induced LVH differs significantly between men and women[124]. In particular, sex-specific LV remodelling leads to a more concentric form of LVH in women. This structural adaptation of LV geometry can be described by less LV dilation and wall thinning in female than male hearts[125]. However, the molecular mechanisms contributing to these sex differences are incompletely understood. So far, analysis of cardiac biopsies from AS patients has revealed a higher level of fibrosis-related genes in male hearts compared to female hearts [126,127]. Studies with experimental animals have also shown higher levels of fibrotic and hypertrophic mediators, as well as lower levels of mitochondrial factors, in male versus female LV tissues under PO[128–132].

However, the investigation of these sex differences has been based on multi-cellular LV tissue samples. Consequently, the use of whole tissues leads to the loss of information on cell-specific gene regulation[133]. The heart consists of a number of different cell types that coordinately regulate cardiac physiology and the response to injury. Given the fundamental role of cardiomyocytes in myocardial function, we aimed here at the assessment of gene expression in isolated human cardiomyocytes. We focused on genes relevant for cardiomyocyte function and adaptation, including hypertrophic markers and genes involved in LV remodelling. We

hypothesized significant sex differences in gene expression, which would be linked to cardiac function in a sex-specific manner.

3.1.1. Cohort characteristics

In this exploratory study, 34 patients (50% men) with AS undergoing aortic valve replacement (AVR) from March 20, 2016, through May 24, 2017 at the German Heart Centre Berlin, were recruited (Tab. 2). Age and body mass index were comparable between men and women. Systolic and diastolic blood pressures were also similar between men and women. There was no statistically significant difference between men and women in the proportion of patients with diabetes, arterial hypertension, concomitant hyperlipidaemia or coronary heart disease. The number of patients treated with calcium antagonists was higher in women compared to men. In contrast, there were more men treated with statins than women. No significant differences were observed between men and women in other relevant medications, such as angiotensin converting enzyme inhibitors, angiotensin II type 1 receptor antagonists, beta-blockers and diuretics.

Echocardiographic assessment of the patients revealed significant sex differences (Tab. 3). The LV end-diastolic diameter was significantly higher in men than in women. Similarly, the posterior wall thickness and LV mass index were higher in men. The EF was significantly lower in men than in women. The aortic valve area index and transvalvular pressure gradients were similar between men and women.

Table 2: **Baseline characteristics of the study population**

	Total (n = 34)	Men (n = 17)	Women (n = 17)	P value
Age, y	68 ± 9	68 ± 10	69 ± 8	0.49
Body mass index, kg/m^2	28 ± 4.9	28 ± 3	28 ± 5.2	0.93
Systolic blood pressure, mmHg	133 ± 24	133 ± 19	133 ± 31	0.98
Diastolic blood pressure, mmHg	72 ± 11	74 ± 10	70 ± 12	0.32
Diabetes Mellitus, %	38	47	24	0.15
Hyperlipidemia, %	41	41	41	1.00
Hypertension, %	73	76	71	0.69
Coronary artery disease, %	47	52	41	0.40
ACE inhibitors, %	42	47	40	0.68
AT1 receptor antagonists, %	28	29	27	0.86
Beta-blockers, %	54	71	40	0.08
Diuretics, %	36	41	33	0.65
Calcium antagonists, %	21	6	40	**0.02**
Statins, %	57	77	40	**0.03**

Values are shown as mean ± SD. ACE, angiotensin converting enzyme; AT1, angiotensin II type 1 receptor

Table 3: **Echocardiographic characteristics of the study population**

	Total (n = 34)	Men (n = 17)	Women (n = 17)	P value
LVEDD, mm	48 ± 7	52 ± 9	45 ± 4	**0.007**
Interventricular septum, mm	14.4 ± 2.8	15.3 ± 3.2	13.6 ± 2.2	0.067
Posterior wall thickness, mm	13.2 ± 2.3	14.2 ± 2.5	12.1 ± 1.6	**0.029**
Left ventricular mass index, g/m^2	140 ± 41	158 ± 45	123 ± 29	**0.037**
Relative wall thickness	0.54 ± 0.1	0.6 ± 0.12	0.5 ± 0.08	0.867
Aortic valve area index, cm^2/m^2	0.4 ± 0.16	0.4 ± 0.19	0.5 ± 0.15	0.962
Mean pressure gradient, mmHg	42 ± 13	39 ± 12	45 ± 14	0.170
Maximal pressure gradient, mmHg	64 ± 18	57 ± 15	69 ± 19	0.066
Left ventricular ejection fraction, %	54 ± 11	49 ± 14	59 ± 5	**0.010**

Values are shown as mean ± SD. LVEDD, left ventricular end-diastolic diameter

3.1.2. Gene regulation in isolated cardiomyocytes of male and female patients

In order to assess cardiomyocyte-specific gene expression between male and female patients, we collected samples from the interventricular septum during AVR and used these to isolate cardiomyocytes. We focused on genes that lead to increased synthesis of structural proteins encoding actin cytoskeleton and cardiac muscle structural and contractile proteins, as well as inflammatory factors. These were ACTC1, CCN2, GATA4, GJA1, MYH6, MYH7, MYL2, MYL3, MYL4, NFKB1, NPPA and NPPB. Our analysis revealed that the levels of the majority of these genes were higher in male cardiomyocytes than in female cardiomyocytes (Fig. 6). Only the levels of MYH7, MYL2 and MYL3 were similar between male and female cardiomyocytes.

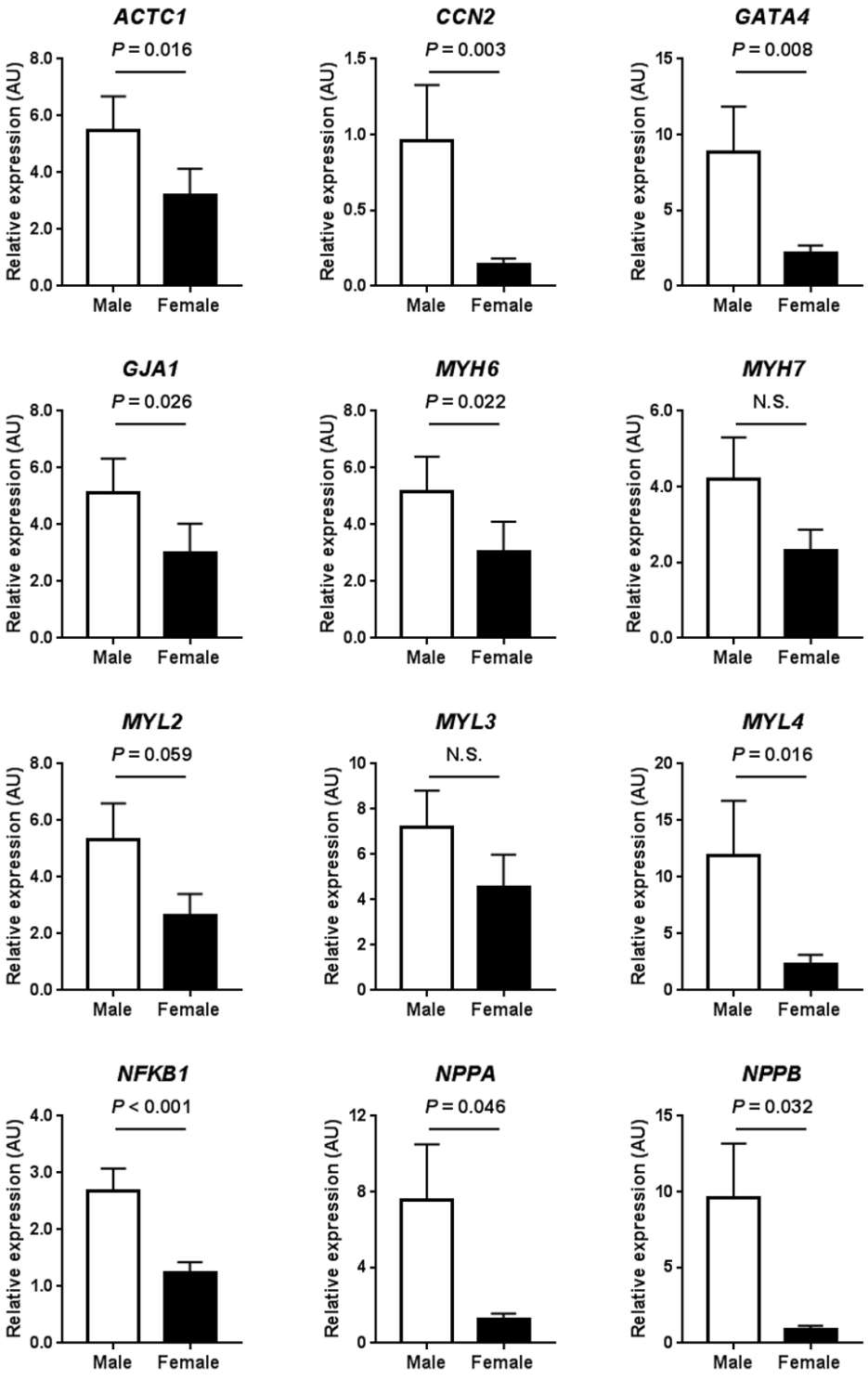

Figure 6: **Gene expression analysis in isolated human cardiomyocytes.** Data present mean ± SEM; n = 12-13/group; AU, arbitrary unit

3.1.3. Relationships between gene expression and ejection fraction

Given the significant sex differences in cardiac function, as assessed by the EF, and in gene expression, we performed a systematic analysis of dependencies between the two. In order to identify relevant links between the EF and specific genes, we combined systematic statistical regression analysis with leave-one-out robustness testing for validation[122]. For an unbiased analysis we considered all potential factors in regression models. In order to avoid over-fitting in view of the available patient numbers, we first focused on the most simple models, each only including one of the 12 genes or one of the clinical features of age, sex, or body mass index. For all of these tests we report a measure of evidence strength (Akaike Information Criterion, AIC) and significance (*p*-value), characterizing the different models in Tab. 4. Specifically, for every patient, we confirmed the relevance of the model features by testing the significance of the model trained on a validation cohort that excludes that patient. We report robustness under two thresholds: 5%, *i.e.* $P < 0.05$, and 10%, *i.e.* $P < 0.1$. We found statistically significant and highly robust gene-based models for the NFKB1 and CCN2 genes (Tab. 4). Our analysis revealed that a higher expression level of NFKB1 was negatively related to EF ($P < 2\%$, robust at $P < 5\%$ in 24/24 leave-one-out validation cohorts). Similarly, a higher expression level of CCN2 was negatively related to EF ($P < 2\%$, robust at $P < 5\%$ in 22/24 and $P < 10\%$ in 24/24 validation cohorts). Notably, a two-fold increase of NFKB1 or CCN2 expression was related to an average reduction of the EF by 11% and 4%, respectively. Considering males and females separately, we identified that the change of EF was significantly different between the sexes ($P < 0.1\%$ for NFKB1 and $P < 1\%$ for CCN2). In particular, a two-fold increase of NFKB1 expression was related to an average reduction of the EF by 17% in males and 7% in female patients (Fig. 7). In the case of CCN2, the average reduction of the EF was 8% and 1% in male and female patients, respectively.

Table 4: **Models with one regressor**

	Beta	SE	*P* value	AIC	robust <5%	robust <10%
ACTC1	-0.03	0.02	0.08	32.46	1/24	8/24
CCN2	**-0.15**	**0.06**	**0.02**	**31.86**	**22/24**	**24/24**
GATA4	-0.02	0.01	0.07	35.08	3/24	15/24
GJA1	-0.04	0.02	0.04	35.22	9/24	21/24
MYH6	-0.04	0.02	0.04	35.36	7/24	20/24
MYH7	-0.06	0.02	0.01	33.80	23/24	23/24
MYL2	-0.04	0.02	0.01	34.92	21/24	23/24
MYL3	-0.02	0.02	0.25	36.98	0/24	2/24
MYL4	-0.01	0.01	0.12	35.83	0/24	6/24
NFKB1	**-0.12**	**0.05**	**0.01**	**29.83**	**24/24**	**24/24**
NPPA	-0.01	0.01	0.09	33.36	3/24	9/24
NPPB	-0.01	0.01	0.55	35.57	0/24	0/24
Age	0.00	0.01	0.68	39.87	0/24	0/24
Sex	**-0.35**	**0.16**	**0.04**	**35.50**	**14/24**	**24/24**
BMI	0.01	0.02	0.39	38.39	0/24	0/24

Beta, model coefficient; SE, standard error of the beta coefficient; P value, statistical

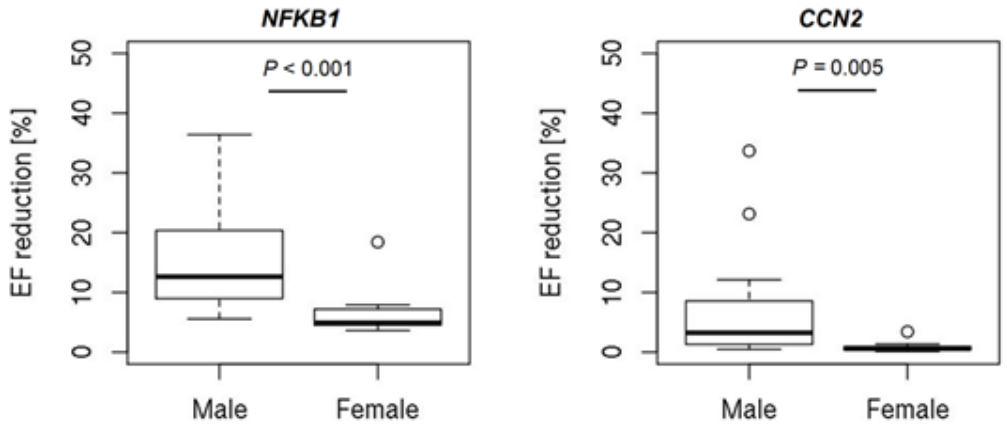

Figure 7: **Sex-specific reduction in the ejection fraction (EF)** upon a two-fold increase in expression of the NFKB1 (left) and CCN2 (right) genes. In each box, the median is presented by the horizontal bolded lines and the corresponding quartiles (whiskers) are shown.

Examining the gene expression for the two robustly significant gene factors suggested that they may be sex-specific. We expressly test this hyptothesis in a model of the sex-specific effects, assessing evidence strength, significance, and robustness as before. Even in this pilot cohort it was possible to identify highly significant sex-specific effects (Tab. 5) that were strong enough to

robustly yield significant models for the two genes, *i.e.* for NFKB1 (P < 2%, robust at P < 5% in 23/24 and P < 10% in 24/24 leave-one-out validation cohorts) and for CCN2 (P < 3%, robust at P < 10% in 24/24 leave-one-out validation cohorts). For both genes, regression extended by the explicit gene-sex interaction term revealed that higher expression was related to a reduced EF in male patients, while there was no significant effect in female patients (Fig. 8). In particular, for NFKB1, a significant male-specific effect was detected (P < 4%, robust at P < 5% in 22/24 and P < 10% in 24/24 leave-one-out validation cohorts), with a two-fold increase yielding an average EF reduction of 15%. For CCN2, a significant male-specific effect was detected (P < 2%, robust at P < 5% in 22/24 and P < 10% in 24/24 leave-one-out validation cohorts), with a two-fold increase being related to an average reduction in EF of over 7%.

Table 5: **Models with interaction terms**

	Beta	SE	*P* value	AIC	robust <5%	robust <10%
***CCN2*:sex**			**0.02**	**33.02**	**22/24**	**24/24**
CCN2:male	-0.13	0.06	**0.02**			
CCN2:female	0.40	0.51	0.41			
***NFKB1*:sex**			**0.02**	**30.62**	**22/24**	**24/24**
NFKB1:male	-0.10	0.05	**0.03**			
NFKB1:female	0.02	0.10	0.81			

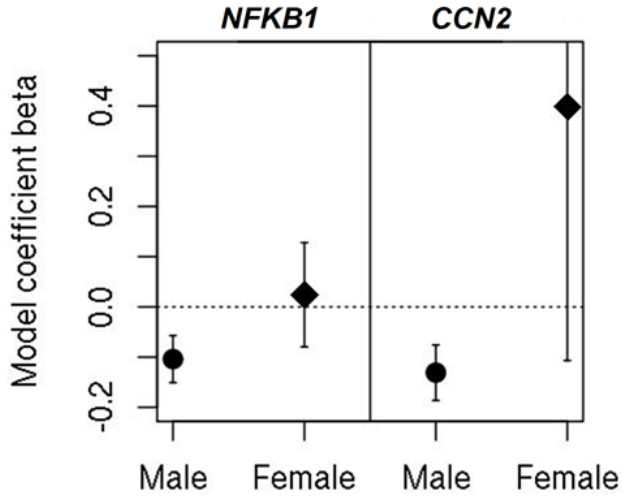

Figure 8: **Beta coefficients of the gene:sex interaction terms.** Bars indicate standard errors for male (circles) and female (diamonds) sex. Notably, the coefficients for male-specific effects have narrow error bars and away from 0, showing their clear significance. In contrast, the female-specific interaction terms are not statistically significant, reflected by them not being distinguishable from 0 (either the coefficient is itself at / close to 0 or the error bars are very broad and crossing 0).

3.1.4. Sex specific differences in the response to pressure overload

The present study is the first to show that in cardiomyocytes isolated from AS patients gene expression differs significantly between the sexes. In particular, the expression of a number of genes associated with maladaptive remodelling was higher in cardiomyocytes of male patients compared to female patients. An additional novel and important finding of this study is that the expression of two inflammatory genes, *i.e.* CCN2 and NFKB1, was negatively related to EF and this was in a sex-specific manner.

The development of PO-induced LVH differs significantly between men and women[125,134,135]. To this extent, we found here sex-specific remodelling of the heart. In particular, at a similar degree of LV outflow obstruction, male hearts had significantly increased LV end-diastolic diameter and decreased EF compared to female hearts. These sex differences in geometry and function may influence outcome. It has been shown that EF and mid-wall stress were better preserved during the progression of AS in women compared to men[136,137]. This indicates that women develop initially a form of remodelling that is more adaptive to PO in terms of function. Of note, although the progression of the stenosis itself was not significantly different between the sexes, women showed 31% lower all-cause mortality independently of treatment, age or blood pressure[137].

Previous studies investigating samples from patients and experimental animals have provided insight into potential mechanisms underlying sex differences in PO[126–132]. However, they were based on multi-cellular LV tissue samples. Consequently, we aimed here at the assessment of gene expression in cardiomyocytes isolated from AS patients. As a pilot study, we took a targeted approach focusing on structural genes, because they are involved in the development of PO-induced LVH and LV remodelling. We found that the expression of ACTC1, GJA1, MYL4, NPPA and NPPB, which are associated with LVH and maladaptive LV remodelling[138], was higher in male cardiomyocytes compared to female cardiomyocytes. In addition, the levels of GATA4, which is an important transcription factor involved in the development of LVH and HF[139], as well as the levels of its downstream targets MYH6 and NPPA, were higher in male cardiomyocytes compared to female cardiomyocytes. Although several factors may influence disease progression and outcome[140,141], the overall induction of these hypertrophic factors in male cardiomyocytes indicates maladaptive remodelling at the molecular level occurring in male patients and supports the view that they may play a key role in the worse outcome observed in men under PO. Notably, the novel finding of increased gene expression of maladaptive remodelling factors in cardiomyocytes of male patients than female patients suggests that the male and female hearts remodel differently through divergent mechanisms. This raises the question to what extent current knowledge, which mostly stems from research on males, is relevant for females, ultimately calling for more research on females.

We further identified that the expression of two inflammation-related genes, CCN2 and NFKB1, was higher in male cardiomyocytes compared to female cardiomyocytes, suggesting a male-specific activation of inflammatory factors. Of note, assessment of the relationship between these

two genes and cardiac function revealed that their levels were negatively related to EF in male patients, pointing at the involvement of inflammatory factors with heart function and failure in a sex-specific manner. Inflammation is a necessary process that ensues stress and injury, such as myocardial infarction and hypertrophy. Consequently, initial activation of inflammatory factors is crucial for the heart to maintain its integrity and function. However, persistent inflammation contributes to disease progression and may lead to HF.

CCN2 encodes cellular communication network factor 2, also known as connective tissue growth factor (CTGF), which is a member of the CCN family of cytokines. Nevertheless, CCN2 does not act as a traditional growth factor or cytokine, but rather as a matricellular protein, induced by transforming growth factor-beta (TGF-β), as well as a cofactor for and downstream mediator of TGF-β[142,143], thereby contributing to cardiac fibrosis[144]. The levels of CCN2 have been previously shown to be induced in cardiac tissues of humans and experimental animals in HF[142,145], but have not been previously studied at an earlier phase of LV remodelling, *i.e.* compensated LVH, or in isolated human cardiomyocytes. In fact, the role of cardiomyocyte-derived CCN2 is not clear, while fibroblast-derived CCN2 appears to be crucial for the development of cardiac fibrosis in mice[146]. We postulate that a pro-inflammatory role might be more relevant for cardiomyocyte-derived CCN2 (discussed below). Collectively, we put forward that the increased expression of CCN2 in male cardiomyocytes under PO may contribute to increased vulnerability to cardiac dysfunction and worse prognosis in men.

NFKB1 codes for the nuclear factor kappa B subunit 1 (NF-κB), which belongs to a family of transcription factors playing a central role in immune and inflammatory responses. Rodent studies have linked NF-κB to the hypertrophic response of the heart. Its prolonged activation is detrimental, triggering chronic inflammation through the induction of inflammatory factors, ultimately leading to HF[147]. Along this line, activation of NF-κB has been reported in cardiac tissues of HF patients[148,149], but as with CCN2, any regulation of NF-κB at an earlier phase of LV remodelling or in isolated human cardiomyocytes is poorly understood. Interestingly, CCN2 has been shown to activate NF-κB in the mouse aorta and kidney, thereby promoting pro-inflammatory factors and inflammatory cell infiltration[150,151]. This suggests a pro-inflammatory role of CCN2 together with NF-κB in male cardiomyocytes. Their persistent induction might render male patients more prone to the development of HF with depressed systolic function.

Our modelling approach revealed that increasing the expression of CCN2 and NFKB1 two-fold yielded a 7% and 15% average reduction in EF of male patients, respectively. This suggests that male-specific activation of pro-inflammatory factors may contribute to LV remodelling and cardiac dysfunction. On the basis of this, we put forward that male patients with compensated LVH might benefit from therapeutic interventions targeting CCN2 and NF-κB. This could address the unmet need for effective anti-inflammatory therapies, since antagonism of the pro-inflammatory cytokine tumour necrosis factor (TNF) in HF has been discouraging[152,153]. It has to be noted, though, that such anti-inflammatory strategies were used in chronic HF. Our data indicate that activation of inflammatory factors already occurs during compensated LVH.

Therefore, we posit that onset of treatment against these targets at an earlier phase of cardiac remodelling may prove useful for the prevention of the progression from compensated LVH to HF. Novel anti-inflammatory strategies targeting the CCN2 and NF-κB pathways deserve further research to identify selective inhibitors reaching therapeutic efficacy and minimizing systemic toxicity.

Moreover, these findings provide insight into the diversity of pathological mechanisms implicated in the development of HF. Over the past years, the clinical definition of HF as a uniform phenotype and pathophysiology has been strongly refuted. Although EF has been used for decades as a clinical quantification of HF, current research has highlighted the high prevalence of HF with preserved EF (HFpEF) and its strong association with the female sex[154,155]. The present findings support the hypothesis that different molecular mechanisms would initiate diverse disease trajectories and highlight the need for more research, in order to understand the background and individuality of every presentation of clinical symptoms of HF, starting with sex-specific differences. Doing so would lead to a more personalized medicine and a better management of this very frequent syndrome with poor outcome.

In conclusion, the present findings show pronounced sex differences in gene expression in isolated human cardiomyocytes and a significant male-specific association between cardiac function and inflammation-related genes. Together, this might contribute to the transition of compensated LVH to HF in the male pressure-overloaded heart, thereby determining postoperative recovery and the development of symptomatic HF with depressed systolic function. Taking these sex differences into account may contribute towards a more accurate design of research and the development of more appropriate therapeutic approaches for both men and women.

Future studies should take into account sex differences, as this will lead to a more accurate design of research and will contribute to the development of more appropriate therapeutic approaches for both men and women.

3.2. Relevance of traditional analyses to clinical research today

The sample size in this study might be considered relatively small. Surely, with a larger sample size, more features could be added to statistical models, without losing interpretability or leading to model over-fitting. However, this cohort is not unusually small, given that cardiac samples were used from patients that have to continue with a functioning heart following biopsy-sampling. This is the case with other patient tissue as well. Blood samples, for example, can be more readily available in larger numbers but will not help us answer the tissue-specific disease-related questions. Notably, the exploratory sample analysed in this study was adequate for identifying significant sex differences in cardiomyocyte gene expression. In fact, this unique pilot

dataset could already show the remarkable strength and robustness of both the overall gene expression effect and the interaction effect with sex, reflected by the solid results obtained.

Such 'Small Data' studies allow for developing simpler models of relations between the few tested variables[122], where complex interactions might show significance, but they do not allow for meaningful interpretation, *e.g.*, possibly overfitting the data. Robustness towards outliers needs to be thoroughly tested in order to exclude the danger of a single patient being responsible alone for the performance of the model. In our analysis we ensure this in the leave-one-out validation. It is also difficult to find biomedical insights in a clinical cohort that will generalize to new independent cohorts[156]. Moreover, qPCR results are known to depend on the choice of housekeeping genes. Specifically, expression data need to be normalized against housekeeping reference genes. Under normal conditions, irrespective of internal and external factors, these reference genes are expected to be expressed in all cells of an organism, because they are essential in maintaining the most basic cellular functions, crucial for the existence of a cell.[157] The commonly used housekeeping genes, however, are not always reliable, with their expression varying under different experimental conditions, possibly adding noise and leading to false results[158–161]. Defining these reference genes is nontrivial and depends on the context. With issues like: deciding between a gene expressed in all tissues, or at a constant level across multiple tissues, with alternative splicing here potentially causing a substantial challenge[157].

4. Chapter 4: Genome-scale analyses

4.1. Big Data enable genome-scale exploration

The apparent limitations of small datasets are overcome when working with genome-scale data, such as gene expression measured by microarrays or RNA-Seq. Housekeeping genes can still be exploited for additional control or the purpose of adjusting unbalanced datasets[47], but are not used for normalization of such expression profiles[162]. On the other hand, such molecular studies are characterized by a very high feature dimensionality, with thousands of genes measured, but again only a relatively limited number of samples, *i.e.* small patient cohorts. This is an issue in various applications of bioinformatics[163–165], sometimes referred to as the *curse of dimensionality*, and it hinders the interpretation of biological signal, with hundreds to thousands of genes found to be implicated in a disease being explored. In order to overcome this issue in an efficient way, it is necessary to reduce the dimensionality of the data with statistical techniques[164,165]. An established way to facilitate analyses of gene expression, by reducing the high dimensionality and the noise inherent to gene expression data[166–168], is a summarization of the expression at pathway level[11]. This is, moreover, a form of data integration, combining prior domain biochemical knowledge with gene expression, and also combining measurements across genes.

Notably, each genome-scale data type has different characteristics, with different type-specific random variation, or noise, present, and bias introduced by multiple different factors[94,166,168–177]. Importantly, there is a stronger correlation of measured effect, like gene expression, between technical replicates than between different subjects, *i.e.* biological replicates[1,178]. In order to distinguish relevant biological signal from technical noise, an appropriate number of biological replicates from each compared condition needs to be accounted for[179]. Apparently, including more biological replicates in an experiment improves the reliability of the identified signal[180]. Thus, naturally, biological variation will be the lowest across replicate cell-line samples, increasing between tissue samples in model organisms, and the highest between human patients for whom usually no technical replicate measurements are available. We analyse the following corresponding datasets in the order of increasing variation.

4.2. Cell line assays as an example for low noise experiments

Biological variation will be the lowest across replicate cell-line samples. This should enable us to perform meaningful analyses with only a small sample count and highly dimensional data. In the following work we employ two well established gene expression technologies each, microarrays and RNA-Seq. Specifically, we design and develop a novel microarray to help optimize Chinese Hamster Ovary (CHO) cells growth, highly relevant for biotechnological processes[181]. We then focus on mouse embryonic stem cells, finding a gene responsible for deriving thyroid progenitors, with important implications for the fields of pluripotent stem cell (PSC) biology and regenerative medicine[2]. In order to reduce the dimensionality of the data and extract higher level biological knowledge we always integrate external knowledge and perform gene set enrichment analyses.

4.2.1. Chinese Hamster Ovary cells profiled by a novel high-performance microarray

Chinese Hamster Ovary (CHO) cells are the most important mammalian cell line[182] for the production of therapeutic proteins due to their ability to perform human-like protein folding with proper glycoforms. They are regarded as safe hosts and can easily be adapted to different culture conditions[183,184]. The adaptation of CHO cells to serum-free and suspension growth is state of the art today, as serum is an ill-defined, variable, animal-derived mixture, which also harbors a contamination risk[185]. Serum-weaned CHO cells will detach and grow in suspension, a beneficial side-effect that allows the switch from static to shaken and stirred cultivation conditions, leading to better scalability. Nevertheless, the adaptation, a labor- and time-intensive process[186], often leads to reduced proliferation, as serum contains many growth supporting factors[185]. Little is known about changes in CHO cell metabolism and gene regulation caused by serum- to protein-free adaptation[187].

Possible approaches to investigate the state of cells include transcriptome, proteome and metabolome analyses. The targeted comparison of cells with distinct characteristics can lead to identification of new engineering targets to enhance cell line performance[188–190]. Vice versa, these technologies are used on engineered cells[191,192] or on cells grown in modified media or under varied culture conditions[187,193,194] to identify deregulated pathways or genes which might be responsible for the observed beneficial cell characteristics. As many of the growth factors present in serum directly interact with the regulation of gene expression, an analysis of the transcriptome appears the most suitable strategy to investigate the effect of serum-weaning on cell behavior.

The transcriptome is usually interrogated using genome-scale profiling platforms such as gene expression microarrays, RNA-Seq, SuperSAGE, or even a combination thereof[195,196].While RNA-Seq has been embraced enthusiastically as a novel method for expression analysis because of its power to discover and profile unknown transcript sequences, work to systematically identify and remove sequence specific measurement distortions is still in early stages[197,198]. In contrast, microarrays have been well established for more than 20 years, and with the development of robust data processing methods they have become a highly effective tool for the routine profiling and comparison of samples. Recent work increasingly suggests that microarrays actually complement RNA-Seq for sensitive genome scale quantitative profiling, with both technologies having rather different strengths and weaknesses[94,166]. In general, RNA-Seq provides better sensitivity and specificity for strongly expressed transcripts, whereas microarrays can obtain a better signal to noise ratio for low expressors, including many transcription factors, which are known to show biological activity also in relatively low abundances[195,198,199]. It has been observed that microarrays more sensitively identify pathways related to protein secretion, which is of particular relevance to biotechnological strain optimization[200]. When target sequences are known, microarrays thus remain a highly attractive tool in the quest to systematically and sensitively assess and analyze the regulation of the transcriptome of CHO cells for optimized biotechnological applications.

Before the first CHO-genome became publically available in 2011[201], either microarrays for Mus musculus were applied for transcriptional profiling of CHO cells [202–205], or CHO-specific microarrays were constructed to target those transcripts identified by expressed sequence tag (EST) sequencing[206–209]. The first CHO-specific microarray contained 2,602 probes for targets mapped by ESTs[208], with later designs including up to 10,000 probes for ESTs[205]. The WyeHamster2a chip was generated by combining ESTs and CHO sequences from public databases, covering a total of 3,567 distinct CHO-specific sequences, representing less than one third of the expressed genes in CHO[203]. These earlier chips left a large part of the transcriptome uncovered, indicating their limited effectiveness for systematic genome scale expression profiling. Recently, full genome chips were generated by collecting RNA-Seq data of several CHO cell lines and Chinese hamster tissues[210] or from a comprehensive CHO transcript database that uses multiple sources of available sequence information[211,212]. Unfortunately, none of the above mentioned designs are publicly available to researchers in the field.

We here therefore introduce and validate a high-performance microarray platform for the scientific community. The efficacy of the platform is demonstrated in an examination of the transcriptome of CHO cells grown in the presence and absence of serum. The obtained molecular understanding of gene expression changes is then successfully used to enhance the growth of serum-free cells by medium optimization.

4.2.1.1. Novel microarray design

A total of 22,036 transcripts were obtained from NCBI, comprising 22,021 unique sequences. Clustering these into groups of highly similar transcripts, led to 20,627 unique transcripts standing alone. In addition, 1,394 sequences formed 620 clusters with over 90% sequence identity, where each cluster is assessed by a joint probe. In all subsequent analyses, we reference groups of clustered sequences by a representative transcript. Thus, the design targets consist of 20,627 singletons and 620 cluster representatives. While a perfect match to all targeted transcripts in a cluster is required and enforced, we aim for minimal cross-hybridization both to other known transcripts, and also potentially unknown transcripts in intergenic regions of the genome. After global probe set optimization for specificity, sensitivity, and uniformity, our design covered 21,565 (98%) of the design targets (Fig. 9), of which 20,853 (95%) were covered with high specificity, showing no detectable cross-hybridization potential to any other transcript or intergenic region, which might harbour unknown transcripts.

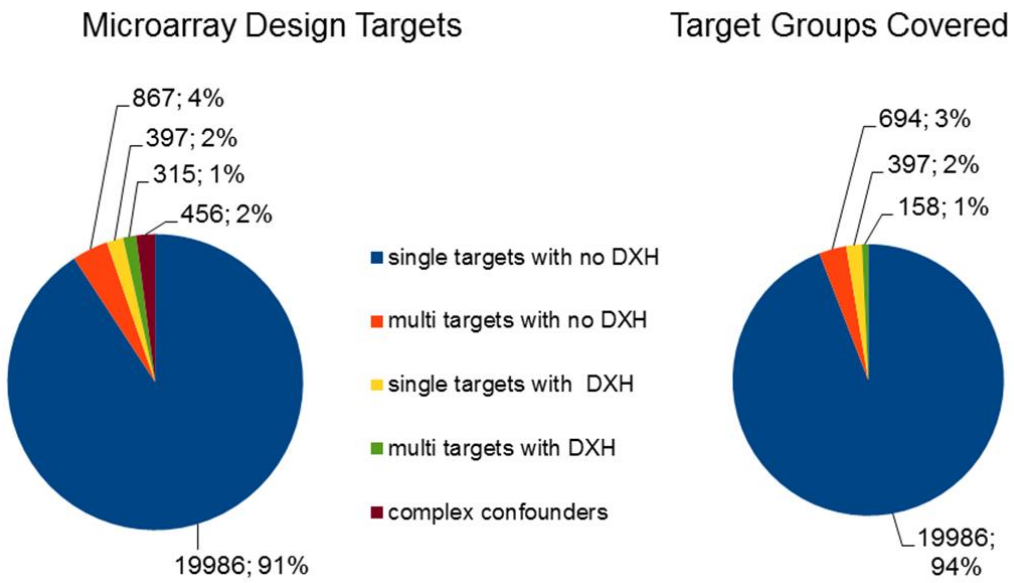

Figure 9: **Summary of design target properties.** The left panel shows the design fate of the 22,021 unique transcripts processed: 91% of them are covered as single targets and with no detectable cross-hybridization (blue), cf. Methods. About 4% fall into multi-target groups of highly similar sequences, such as arising from a recent gene duplications, which can be measured together showing no detectable cross-hybridization (orange), for a total of 95% that can be assessed with high-specificity. For single targets and clusters of highly similar targets ('multi targets'), respectively, only 2% and 1% could not be assessed without detectable cross-hybridization (yellow and green). Finally, just 2% of all design targets showed complex interaction patterns so that conservative probes could not be

obtained (brown). The right panel shows a summary of properties for the target groups covered by the microarray design. The same colour code is used. 97% of all target groups can be assessed with high-specificity. In the legend box, 'detectable cross-hybridization' is marked as 'DXH'.

We then confirmed the robustness of our design, with regard to a major new annotation release, with the original design retaining specificity and comprehensive coverage even after this substantial annotation update. Furthermore, we mapped all the specific probes to the new annotation to estimate the genes covered by probes of our design, demonstrating stringent specificity of probes and ensuring that designs will remain robust also under subsequent annotation updates[181].

4.2.1.2. Media supplement optimization through pathway information

RNA samples of CHO-K1 cells, grown adherent in medium containing 5% fetal calf serum, were compared to samples obtained from the same CHO-K1 cell line after adaptation to growth in suspension in a protein-free medium[181]. Based on the prior understanding of the response to serum withdrawal, one would expect these samples to have an altered transcription pattern reflecting (a) the removal of growth factors and hormones as well as small molecules present in serum or bound to serum-components from the medium and (b) the morphological, mostly extracellular matrix related changes that occur with adaptation to growth in suspension. We find 369 genes differentially expressed under a *q*-value of 0.05[181].

GO[213,214] annotation from mouse was transferred to CHO by predicted orthology because no systematic GO assignments were yet directly available for CHO. Significantly enriched GO-terms are shown in Tab. 6, and the graph indicating their positions in the respective GO trees is provided in Fig. 10. Out of the 120 GO terms with *q*-values $\leq$ 0.05, 54 were related to extracellular matrix and 4 were related to lipid metabolism (Tab. 6). The other terms were connected to cellular response to the environment and to metabolism.

Table 6: **Enriched Gene Ontologies**. GO terms related to extracellular matrix and lipid metabolism are highlighted. Cellular component, biological process and molecular function are abbreviated as “C”, “B” and “M” respectively.

ID	p.adjusted	GO category	name
GO:0044421	2.41E-07	C	extracellular region part
GO:0005576	2.41E-07	C	extracellular region
GO:1901615	7.03E-05	B	organic hydroxy compound metabolic process
GO:1901617	2.91E-04	B	organic hydroxy compound biosynthetic process
GO:0008299	0.001328479	B	isoprenoid biosynthetic process
GO:0008202	0.001356965	B	steroid metabolic process
GO:0006694	0.009243104	B	steroid biosynthetic process
GO:0006066	0.016104007	B	alcohol metabolic process
GO:0065010	0.016431729	C	extracellular membrane-bounded organelle
GO:0006720	0.027220461	B	isoprenoid metabolic process
GO:0044710	0.036249463	B	single-organism metabolic process
GO:0043230	0.036249463	C	extracellular organelle
GO:0006950	0.036249463	B	response to stress
GO:1901681	0.036249463	M	sulfur compound binding
GO:0010035	0.036249463	B	response to inorganic substance
GO:0005539	0.036249463	M	glycosaminoglycan binding
GO:0008152	0.036249463	B	metabolic process
GO:0003823	0.036249463	M	antigen binding
GO:0031012	0.042498891	C	extracellular matrix
GO:0016765	0.044196523	M	transferase activity, transferring alkyl or aryl (other than methyl) groups

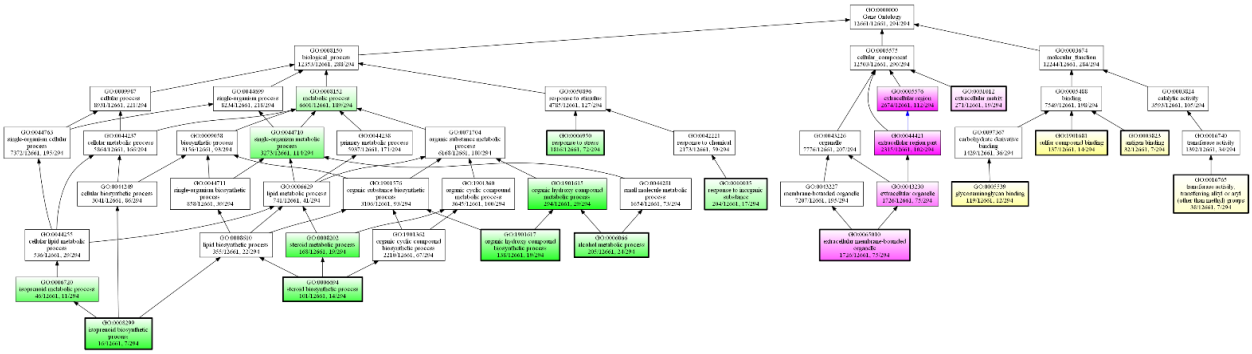

Figure 10: **GO-Tree**. Enriched GO-terms due to serum-free adaptation are depicted in green (biological process), purple (cellular component) and yellow (molecular function).

We then analysed enriched KEGG[215] pathways. The 369 differentially expressed transcripts of our novel chip, which corresponded to 364 genes, were mapped to the KEGG pathways. 170 of the differentially expressed genes were associated with one or more pathways. 10 pathways were highlighted as significantly enriched (q-value ≤ 0.05; Tab. 7) and the genes were highlighted on these maps indicating their fold change[181]. Of these, 3 were related to lipid and fatty acid metabolism (KEGG-ID: 00072, 00100 and 00900). Glutathione metabolism is related to antioxidant defense and detoxification. The other terms (KEGG-ID: 00980, 00982, 01524, 05204) were related to cytochrome p450 and cancer, and were significant because of their high

number of glutathione-related genes, which account for 75 – 80 % of the regulated genes in these pathways. Also, glutamine- and glutamate metabolism, and in general metabolic pathways were indicated as differentially regulated.

Table 7: **Enriched KEGG-pathways**

KEGG ID	name	genes in pathway	total genes in pathway	proportion	p.adjusted
72	Synthesis and degradation of ketone bodies	4	12	0.333	1.61E-03
100	Steroid biosynthesis	7	20	0.35	6.40E-06
471	D-Glutamine and D-glutamate metabolism	2	3	0.667	3.28E-03
480	Glutathione metabolism	11	51	0.216	2.25E-06
900	Terpenoid backbone biosynthesis	8	23	0.348	2.25E-06
980	Metabolism of xenobiotics by cytochrome P450	11	72	0.153	5.51E-05
982	Drug metabolism – cytochrome P450	11	60	0.183	8.93E-06
1100	Metabolic pathways	45	1152	0.039	5.56E-03
1524	Platinum drug resistance	13	75	0.173	2.25E-06
5204	Chemical carcinogenesis	12	81	0.148	3.04E-05

As lipid metabolism-related GO[213,214] and KEGG terms were the most prominently differentially regulated pathways, we tested a standard lipid formulation supplement for its effect on growth rate in protein-free medium. Although no GO and KEGG pathway enrichment was observed for nucleotide metabolism, several genes involved in this process were differentially regulated, including two that are known to be required for nucleotide synthesis (Hprt1, Atic). As faster growing cells can be expected to have a higher requirement for nucleotides, we also tested a nucleotide precursor supplement consisting of hypoxanthine and thymidine (HT).

Both supplements were used at two different concentrations, alone and in combination. After an initial adaptation phase of two weeks, two batches were performed consecutively, each consisting of two replicates for each condition. While each supplement alone showed no (lipid supplement) to minor (HT supplement) growth enhancement, the combination of both lead to an increase in growth of 10% (Fig. 11A). The growth rate was lower when calculated via viable cell volumes, as the viable cell size decreases from day 0 to day 3 (see Fig. 11B). Another effect observed was the reduction in viable cell size induced by supplementing the cells with hypoxanthine and thymidine (Fig. 11B). This effect was most prominent on day 1 (20% size reduction), and vanished again on day 4 (HT 1:100 supplemented) or day 5 (HT 1:50 supplemented). Thus, it seemed to be highly dependent on the availability of HT supplement in the medium and was quickly reversible.

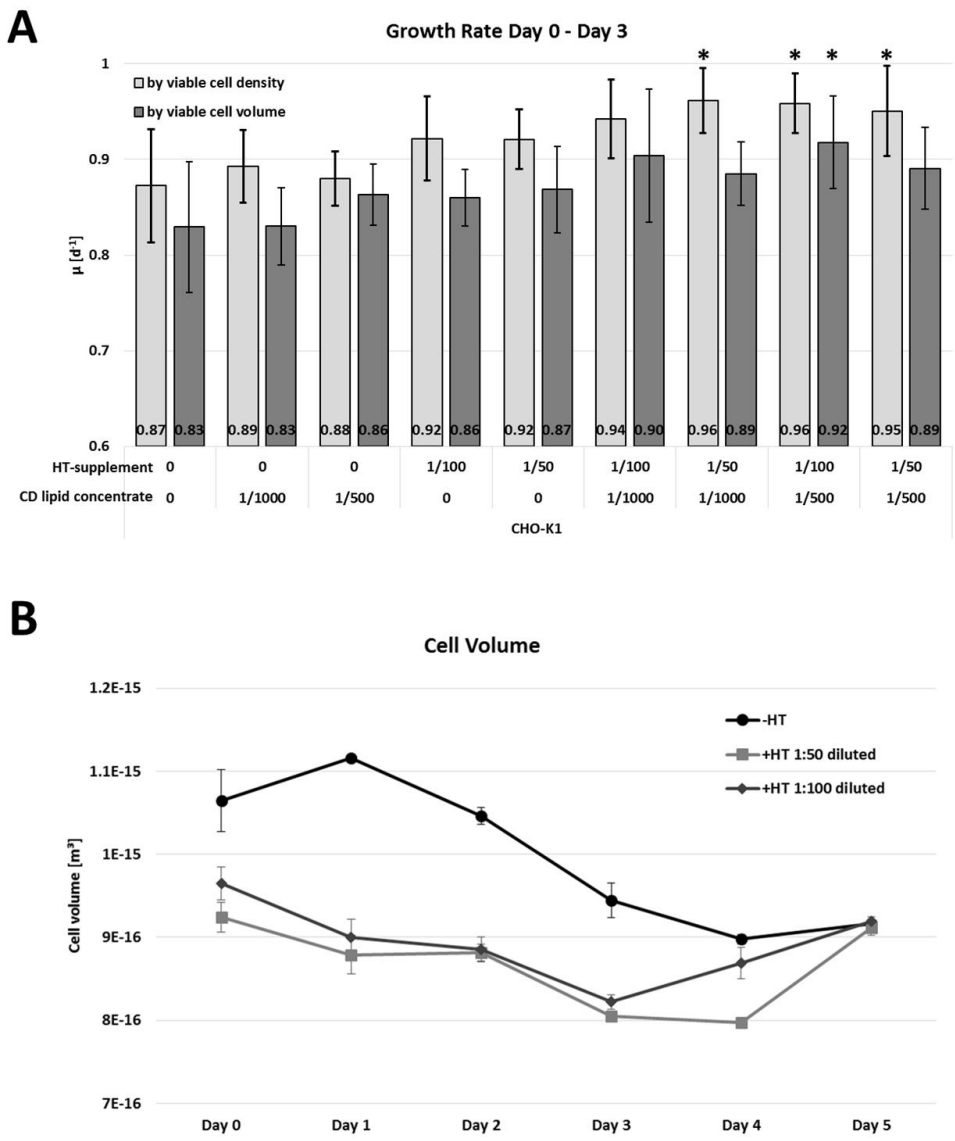

Figure 11: **Media supplement batches.** A: Growth rate [μ] in exponential growth phase (72h). Stars indicate significant changes in growth when compared to the cells without media supplements. B: Cell volume grouped by the amount of HT-supplement added. The –HT line includes all samples without nucleotide addition, with or without addition of lipids. The bars/points represent the mean ± standard deviation of two batches, each consisting of two replicates.

4.2.1.3. Enhancing cell growth via detection of transcriptomic differences

The adaptation of CHO-K1 cells to serum-free media in general leads to reduced growth in suspension. To understand the phenotypic changes caused by serum deprivation, we compared the gene expression pattern of cells cultivated in the presence of serum and of cells adapted to growth in suspension in protein-free medium using microarrays. To this end, we have introduced and validated the first public CHO-specific microarray, using state-of-the-art model based probe design. Highly specific probes with no detectable cross-hybridization were designed for 95% of

the targets, with 93% comprising single transcripts, and the remainder detecting groups of highly similar transcripts (Fig. 9). It is noteworthy that transcripts with high similarity also cannot be quantified reliably using RNA-Seq, because only a small fraction of reads will hit sequence regions discriminating similar sequences, and the resulting high noise is particularly an issue for complex transcriptomes, such as those of mammals[198]. The main power of next-generation sequencing methods lies in sequence discovery, mapping out the transcriptome of different species. Indeed, for effective quantitative expression profiling by RNA-Seq, data analysis has to exploit known gene models. Conversely, with custom microarray platforms well established, once a revised genomic sequence or new gene modes become available from sequencing data, microarray designs can be adapted on the fly to immediately take advantage of the latest updates, with probes chosen to directly target the most informative sequence regions for effective quantitative profiling. We have moreover shown that the performance of the platform design was robust even to substantial transcript annotation updates.

We have applied this array to understand the biology involved in CHO adaptation to protein-free medium and suspension growth. As expected, after GO and KEGG pathway enrichment analysis several GO-terms were found to be enriched that are linked to the extracellular matrix, which is remodelled due to transition from adherent to suspension growth (Tab. 6). In addition, significant changes in sensing of the environment and in metabolism were detected (see Tab. 6, Tab. 7). There was significant impact on isoprenoid and several lipid biosynthesis pathways, which were downregulated in the serum-supplied culture. This could be due to the fact that serum contains significant amounts of steroids and cholesterol, bound to carrier proteins such as albumin and lipoproteins, thus provided in a soluble form[185], while protein-free media provide these substances only in small amounts due to their low solubility. CHO cells are able to synthesize their own isoprenoids and lipids, and the amounts required to be synthesized will depend on the availability in the medium[216]. These changes can also be explained by the necessity to change the molecular composition of the plasma membrane for cells growing in suspension rather than as adherent cells[217,218].

The terms antigen-binding, response to stress and glutathione metabolism were also significantly enriched. Most of the associated genes were downregulated in the serum-supplemented culture, indicating that the CHO-K1 cells were under less stress than the protein-free suspension adapted cells, as stress induces genes connected to glutathione metabolism[219,220] and antigen-presentation[221,222]. This is probably caused by the mechanical force and the higher oxygen supply which was applied to the serum-free and suspension-adapted CHO-K1 cells due to the dynamic cultivation system used[223] and the higher cell density per ml culture volume. Though the regulation of growth-factor related signaling cascades is highly associated with phosphorylation rather than with transcriptomic changes[212], it is still surprising that complete removal of all growth factors caused no significantly regulated pathway.

Still, the GO and KEGG pathway enrichment analysis can only find a portion of all phenotypical relevant transcriptomic changes, considering that not all differentially expressed genes are

annotated with GO-terms or KEGG-IDs and that small number of significantly regulated genes are often not detected by enrichment analysis. Thus, the differentially expressed gene list[181] complements the information extracted by the enrichment analyses. Several genes involved in nucleotide metabolism, with two of them (Hprt1, Atic) associated to nucleotide synthesis, were found to be differentially regulated, though no GO and pathway enrichment was indicated.

Considering the results reported above, we next aimed to enhance the phenotype of serum-free and suspension adapted CHO-K1 cells by media supplementation. As lipid biosynthesis includes many enzymatic steps and requires ATP as an energy source, we decided to supplement the cells with cholesterol and other lipids to unburden this biosynthetic machinery. In addition, based on the observed upregulation of nucleotide synthesis in the protein free culture, nucleotides were added to the media. With a combination of these supplements, a significant growth enhancement could be induced (Fig. 11A, $p < 0.05$, t-test). The energy which had been used for lipid biosynthesis has now become available for other cellular processes. Also the supplementation of cells with hypoxanthine and thymidine should shift the cells more from de novo synthesis to salvage pathways, which consume less energy. When HT-supplement was added to the cells, a reduction in cell size occurred in addition to growth enhancement. This seemed to be reversed when the supplement is depleted (Fig. 11B). An increase in maximum cell density due to HT supplementation which was reported in a previous study[224] could not be observed here. As HT-supplemented cells were already smaller when seeded due to the adaptation phase and as the relative change of cell size reduction over the batch was similar for cells with and without HT-supplement, the measured growth enhancement by supplementation was not affected by the cell size reduction. Further investigation is necessary to determine the detailed characteristics of the phenotype induced by the HT-supplement.

With the first public genome-scale CHO-specific chip design introduced and validated here, we were able to sensitively detect transcriptomic differences which occur in CHO-K1 cells due to adaptation to protein-free and suspension growth. Based on the results, it was possible to design a medium supplementation strategy which enhanced the growth phenotype. Additionally, an effect on the cell size was observed when HT was added to the media.

4.2.2. Mouse embryonic stem cells for deriving thyroid progenitors via gene overexpression

Tissue progenitors are important intermediate cell types in embryonic development. Since these cells give rise to all mature cell types within a given tissue, their in vitro derivation has important implications for the fields of pluripotent stem cell (PSC) biology and regenerative medicine. Significant advances in anterior foregut endoderm (AFE) progenitor biology in recent years[225] have led to derivation of AFE lung and thyroid progenitors (TPs) and their clinically relevant progeny from PSCs[226–232]. However, with one notable exception[227], efficiencies of progenitor derivation have been relatively low (<40%).

Overexpression of transcription factors (TFs) is a well-established approach to manipulate cellular identities as it results in reconfiguration or emergence of core TF networks, with derivation of induced pluripotent stem cells (iPSCs) from somatic cells being the most prominent example[233]. Inducible TF expression in PSCs has been primarily used to enhance the derivation of differentiated progeny[234–237]. For example, Costagliola and coworkers[234] used forced overexpression of the thyroid TFs *Nkx2-1*/*Pax8* in mouse embryonic stem cells (mESCs) to derive thyrocyte-like cells with high efficiency (60%) that formed in vitro follicular structures and rescued athyroid mice upon transplantation. Despite the impressive findings, this work neither focused on developmental stage competence nor examined the mechanisms of this conversion. Overall, studies that systematically investigate the interplay between pulsed heterologous TF expression and developmental stages in directed differentiation of PSCs are lacking. Such studies may provide valuable insights in the cellular plasticity of developmental intermediates and lead to rational design of robust and efficient directed differentiation protocols. To address these questions, we used thyroid directed differentiation[228] in combination with transient *Nkx2-1* overexpression as our model system.

We make use of novel computational methods[238,239] and genome-wide gene expression analysis by RNA-Seq to demonstrate that the thyroid-like cells induced by NKX2-1 are similar to mouse embryonic thyroid epithelial cells. Finally, we employ mathematical modeling to suggest that the *Nkx2-1* overexpression effect is potentially governed by a time-dependent bistable switch[2].

4.2.2.1. Stem cell systems model complex disease states

PSC-based systems hold great promise for the mass production of transplantable, clinically relevant cell types and for *in vitro* modeling of complex disease states. Currently, a major roadblock to achieving these goals is the poor or variable differentiation efficiency of many differentiation protocols. This study used developmental-stage-specific overexpression of a single TF, NKX2-1, to (1) investigate how cell competence changes in a dynamic, developmentally relevant system and (2) improve the efficiency of thyroid progenitor specification.

Our results demonstrate that the inductive effect of NKX2-1 is restricted to a singular stage of competence contingent on several synergistic parameters (FOXA2 levels, duration of anteriorization, and BMP4/FGF2 signaling), implying that AFE-staged cells possess a unique epigenetic status allowing for robust conversion to thyroid (Fig. 12). Respiratory lineages were not derived as indicated by insignificant numbers of SPC+ cells and the only presumed lung marker with high levels of expression (Sftpb) is expressed in both adult mouse (Figure S2A in Dame *et al*[2]) and human thyroid (http://gtexportal.org/home/gene/SFTPB). We presume that the absence of Wnt agonists, an important signal for lung specification[227,240,241], is the major reason for the paucity of respiratory lineages following NKX2-1 overexpression in our system.

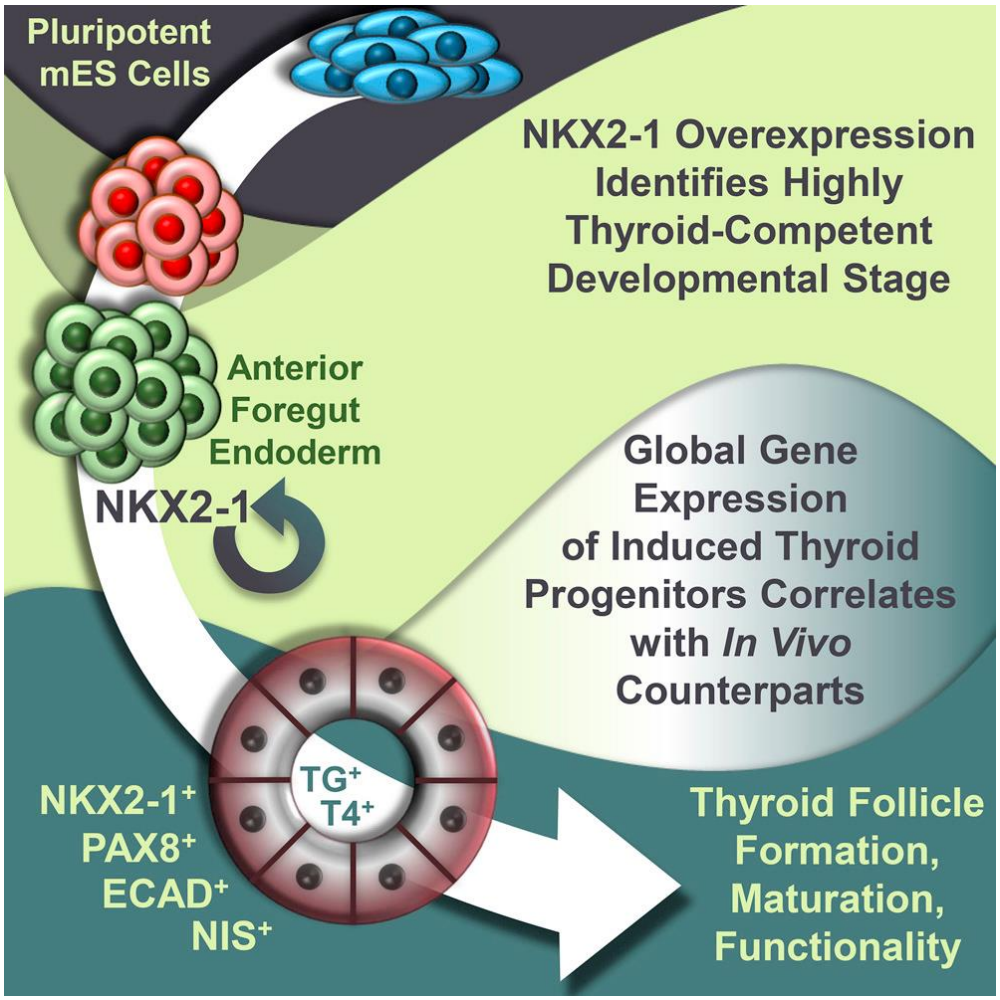

Figure 12: **Graphical depiction of the developed differentiation system**. Anterior foregut endoderm (AFE) is robustly differentiated to thyroid progenitors via overexpression of NKX2-1**.**

4.2.2.2. *In vitro* might be on par with *in vivo* models in some applications

Our study also achieved the practical goal of highly efficient thyroid differentiation. Our data indicate that the majority of the derived progenitors acquired a thyroid identity comparable to their *in vivo* counterparts, and importantly, their progeny gave rise to follicular-like structures in a 3D environment, expressed genes of thyroid hormone biosynthesis at levels comparable with adult thyroid, and produced high levels of T4 hormone. Overall, it appears that brief NKX2-1 exogenous expression during a well-defined window of maximum thyroid competence is sufficient to dramatically increase the yield and robustness of PSC thyroid specification and differentiation. Although a similar end-stage result has been previously reported through direct reprogramming of mESCs[234], our approach delves into the mechanistic aspects of thyroid fate decisions and competence at a developmentally relevant stage. Further work is needed to define whether the thyrocyte-like cells produced downstream in our system are functionally equivalent to the reprogrammed cells and to the progeny of purified bona fide TPs[228].

4.2.2.3. NKX2-1 overexpression leads to efficient thyroid derivation

We report that NKX2-1 can function as an inductive signal at the AFE stage of thyroid directed differentiation (Fig. 12) as its transient expression converts AFE-staged cells to thyroid-like cells with high efficiency (>70%)[2]. The resulting cells differentiate to thyrocyte-like cells that self-organize to epithelial, follicle-like structures in 3D Matrigel culture. The effect is tightly regulated and pertains only to a narrow developmental window of competence defined by and contingent on dual BMP and FGF signaling, thyroid-competent AFE subpopulations and correct anterior patterning of definitive endoderm.

This system can also be useful as a discovery tool in the absence of lineage-specific reporters. Analysis of the RNA-seq has identified cell surface markers as well as potential regulators of thyroid fate in AFE and early thyroid progenitors. Some of the TFs identified (Irx5, Hoxb8, Isl1) have been involved in mouse foregut and thyroid development[242,243], while others, such as PROX1, have been implicated in human thyroid disease[244]. Future PSC-based and in vivo studies will unravel the thyroid-related function of select TFs during thyroid specification and development.

Our theoretical model proposes a bistable positive feedback loop as the underlying mechanism to describe the endogenous Nkx2-1 locus activation at the AFE stage leading to subsequent stabilization of the core TF network, establishing thyroid identity. Future work will focus on experimental validation of the model and investigate the possibility that bistable switches controlling bifurcation dynamics in cell-fate decisions[245] can lead to the development of highly efficient and robust protocols of general applicability.

4.3. Tissue samples from model organisms

We next look at two different datasets of model organisms that should reflect human biology even better, RNA-Seq expression of mouse[99], and proteomics of sheep[1]. Model organisms, like the widely researched mouse, allow scientists to extensively study evolutionary conserved biological processes and developmental mechanisms, with discoveries made potentially explaining also human biology[246,247]. In particular, model organisms should, however, also allow us to study human disease, where experimentation on human subjects would be impossible and unethical. Even though mice are widely used in research, they differ physiologically from humans to an extent that makes them less desirable as models of human disease, with the links between genes and disease most probably being different than in humans[246,248]. Physiologically more similar to humans in both size and longevity, larger animals, like sheep, should provide a better alternative to mice for studying specific human condition[248].

4.3.1. Mouse model for finding the genetic program of differentiation to lungs

Pluripotent stem cell (PSC)-based systems offer the possibility of de novo somatic cell derivation through directed differentiation, a multistage process with recapitulation of developmental milestones[249]. This methodology relies heavily on prior knowledge of developmental pathways and processes within the tissue/organ of interest and incomplete developmental knowledge can be a significant impediment to the establishment of efficient directed differentiation protocols.

This is well-illustrated by the late emergence of protocols for anterior foregut derivatives, such as lung, thyroid and thymus as opposed to the more mature research areas of pancreatic/hepatic differentiation[225]. In particular, the impressive progress in the last ten years in PSC-derivation and downstream differentiation of respiratory and thyroid lineages coincided with major advances in understanding of cell fate decisions in vivo. In the case of thyroid specification, *in vitro* findings were subsequently validated *in vivo*[228]. In the case of lung specification[228], elucidation of the role of Bmp and Wnt signals through loss-of-function studies in murine development[240,241,250] has led to the development of growth factor cocktails for derivation of lung progenitors from mouse and human PSCs[226,227,229,251]. Similarly, the in vitro derivation of ciliated cells[252–254] as well as proximal and distal lung progenitors[255,256] were made possible by in vivo studies of Notch signaling in ciliated cell differentiation[257] and Wnt signaling in lung proximal-distal patterning[258–260], respectively.

An important intermediate population in lung lineage derivation from PSCs is the lung epithelial primordial progenitor population, *i.e.* the first cells within the foregut that are specified to a lung epithelial cell fate, but do not yet express markers of lung epithelial differentiation, such as surfactant genes. *Nkx2-1*, the earliest marker of lung fate, is first expressed in the prospective lung domain of the gut tube at around embryonic (E) day 9.0 during mouse embryogenesis[261]. It is also expressed in the prospective thyroid domain, specified at around E8.5, and in the developing forebrain[262]. Previous work with mouse and human PSC lines, from our group and others, has therefore used *in vitro* Nkx2-1 expression as a marker of lung progenitors[227,229,230,251,255,256,263] with both distal alveolar and proximal airway epithelial competence. Additionally, our previous work using Nkx2-1 fluorescent reporter lines has allowed us to purify and study endodermal Nkx2-1$^+$ populations[228,229,251,263]. Nevertheless, in the absence of similar characterization of the earliest primary Nkx2-1$^+$ progenitors that arise during mammalian organogenesis, a major question remains, how closely do the *in vitro* progenitors resemble their *in vivo* counterparts? Indeed, previous efforts to study early thyroid and lung patterning have used microdissection techniques that inevitably lead to isolation of mixed populations[242,264]. Furthermore, the equivalent stage in human development is not readily accessible and *in vivo* information on the lung primordium, even Pluripotent stem cell (PSC)-based systems offer the possibility of de novo somatic cell derivation through directed differentiation, a multistage process with recapitulation of developmental milestones[249]. This methodology relies heavily on prior knowledge of developmental pathways and processes within

the tissue/organ of interest and incomplete developmental knowledge can be a significant impediment to the establishment of efficient directed differentiation protocols.

This is well-illustrated by the late emergence of protocols for anterior foregut derivatives, such as lung, thyroid and thymus as opposed to the more mature research areas of pancreatic/hepatic differentiation[225]. In particular, the impressive progress in the last ten years in PSC-derivation and downstream differentiation of respiratory and thyroid lineages coincided with major advances in understanding of cell fate decisions *in vivo*. In the case of thyroid specification, *in vitro* findings were subsequently validated *in vivo*[228]. In the case of lung specification[228], elucidation of the role of Bmp and Wnt signals through loss-of-function studies in murine development[240,241,250] has led to the development of growth factor cocktails for derivation of lung progenitors from mouse and human PSCs[226,227,229,230,251]. Similarly, the in vitro derivation of ciliated cells[252–254] as well as proximal and distal lung progenitors[255,256] were made possible by *in vivo* studies of Notch signaling in ciliated cell differentiation[257] and Wnt signaling in lung proximal-distal patterning[258–260], respectively.

An important intermediate population in lung lineage derivation from PSCs is the lung epithelial primordial progenitor population, *i.e.* the first cells within the foregut that are specified to a lung epithelial cell fate, but do not yet express markers of lung epithelial differentiation, such as surfactant genes. *Nkx2-1*, the earliest marker of lung fate, is first expressed in the prospective lung domain of the gut tube at around embryonic (E) day 9.0 during mouse embryogenesis[261]. It is also expressed in the prospective thyroid domain, specified at around E8.5, and in the developing forebrain[262]. Previous work with mouse and human PSC lines, from our group and others, has therefore used *in vitro* Nkx2-1 expression as a marker of lung progenitors[227,229,230,251,255,256,263] with both distal alveolar and proximal airway epithelial competence. Additionally, our previous work using Nkx2-1 fluorescent reporter lines has allowed us to purify and study endodermal Nkx2-1$^+$ populations[228,229,251,263]. Nevertheless, in the absence of similar characterization of the earliest primary Nkx2-1$^+$ progenitors that arise during mammalian organogenesis, a major question remains, how closely do the *in vitro* progenitors resemble their *in vivo* counterparts? Indeed, previous efforts to study early thyroid and lung patterning have used microdissection techniques that inevitably lead to isolation of mixed populations[242,264]. Furthermore, the equivalent stage in human development is not readily accessible and *in vivo* information on the lung primordium, even from other lunged species, would be highly informative for human PSC-based lung differentiation studies and future regenerative medicine applications.

Here, we report the identification, purification, and global transcriptomic analysis of the earliest detectable lung progenitors that compose the developing mouse lung primordium *in vivo*. We compare the genetic program of these developing lung progenitors to foregut endodermal precursors purified just prior to the onset of expression of the Nkx2-1 program and contrast these programs to those of the Nkx2-1$^+$ lineages that compose the developing forebrain and thyroid.

These comparisons delineate the unique genetic program of the lung primordium at the moment of lung lineage specification *in vivo* and suggest pathways that regulate the emergence of lung fate within the developing foregut endoderm. We then mine the global transcriptome of the *in vivo* lung primordium to improve the critical stage of *in vitro* lung specification in a mouse PSC-based system. We show that modification of cell culture conditions, including biomechanical properties of the culture substratum can enhance the epithelial character of PSC-derived mouse lung progenitors. By applying computational methods for classifying cell fate consisting of Linear Algebra-based projections, we demonstrate the improved fidelity of progenitors derived under novel conditions as reflected by increased projection scores onto the *in vivo* lung epithelial primordial population. Thus, our approach delineates the unique genetic program of the earliest primordial lung progenitors and underscores the importance of applying such an *in vivo* "roadmap" to guide the specification of primordial organ progenitors that may be engineered *in vitro* from alternate sources, such as PSCs.

4.3.1.1. Finding distinct genetic program of lung through transcriptomic analysis

Directed differentiation of PSCs has emerged as one of the most promising regenerative medicine platforms, for disease modeling and future cell replacement therapies[265]. One of the main remaining obstacles in the clinical use of PSCs for the treatment of lung disease is the partial understanding of the sequence of cell fate decisions that lead from PSC-derived anterior foregut endoderm to functional, clinically-relevant lung cell populations. *In vivo*, the formation of these cell types and, in fact, of all lung epithelial lineages depends on the critical moment of lung specification, *i.e.* the decision of specific cells within the anterior part of the gut tube to become lung primordial progenitors. It is then evident that an incomplete understanding of this particular progenitor population can limit attempts to generate a similar population *in vitro*. By providing global transcriptomic signatures of the earliest in vivo Nkx2-1^{+} progenitors and quantitative algorithms for their application, our work establishes important benchmarks against which cells engineered *in vitro* can be compared as they proceed through the developmental gateway of lineage specification. Consequently, our current work addresses deficiencies in the development of lung differentiation protocols by applying computational methods to improve the efficiency of *in vitro* lung progenitor derivation.

The underlying assumption of current lung directed differentiation protocols is that lung epithelial lineages *in vivo* are derived through an Nkx2-1^{+} intermediate primordial progenitor[226,227,230,263,266]. We substantiated this statement by performing indelible marking of Nkx2-1^{+} lung progenitors around E9.0 using an inducible Nkx2-1 knock-in Cre driver and an nT/nG reporter mouse[99]. Nuclear GFP-positive cells were exclusively found in the epithelial (EPCAM^{+} or CDH1^{+}) compartment of the developing lung at both E14.5 and E18.5 and most importantly in cells expressing markers of ciliated, secretory, basal, AEC1 and AEC2 lung epithelial lineages. These findings further support the wide use of Nkx2-1 as the earliest marker of lung epithelial fate and strongly imply that the use of Nkx2-1 fluorescent reporters in mouse

and human PSCs[229,263] in combination with lung specification cocktails leads to derivation and isolation of putative lung progenitors with broad differentiation competence.

Using our Nkx2-1^{GFP} knock-in mouse we were able to sort Nkx2-1^{+} lung epithelial progenitors at E9.0 and characterized their transcriptome by bulk RNA-Seq (Fig. 13).

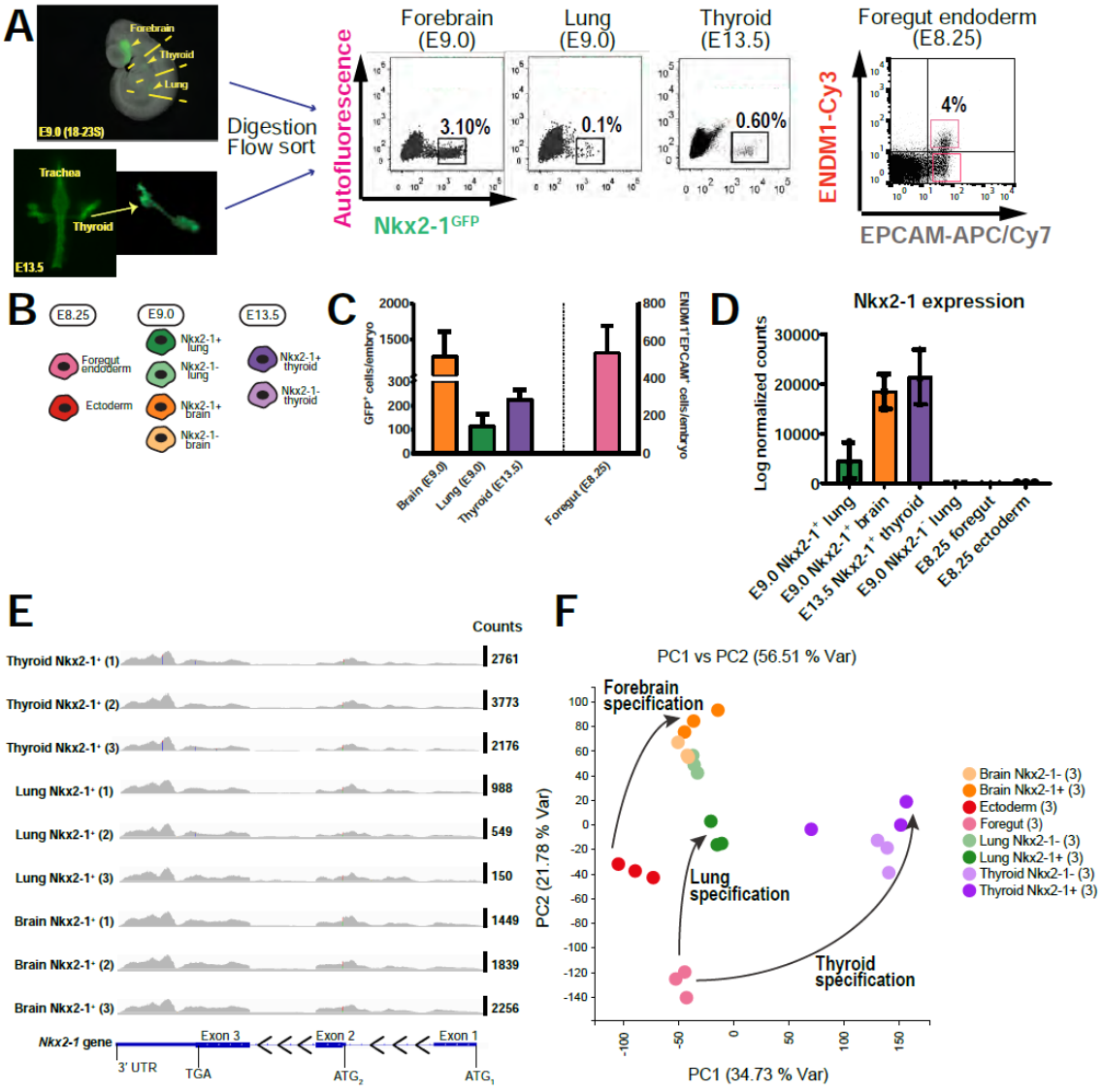

Figure 13: **RNA-Seq analysis of purified Nkx2-1^{+} endodermal and ectodermal cell populations during mouse development.**

(A) Schematic of embryo dissection and NKX2-1^{GFP+} cell sorting at the lung primordium stage (E9.0, 18-23 somites) and at E13.5. The Nkx2-1^{GFP+} lung, thyroid and forebrain domains were micro-dissected using an epifluorescence stereomicroscope. At E13.5, thyroid is separated from the trachea prior to enzyme digestion and sorting (left panels). Bivariate flow cytometry dot plots showing sorted NKX2-1^{GFP+} cell populations (middle panel) and pre-specified foregut endoderm (ENDM1^{+}EPCAM^{+}) and ectoderm (ENDM1^{-}EPCAM^{+}) (right panel).

(B) FACS-purified cell populations used in RNA-Seq analysis. The same colors are consistently used in subsequent figures to identify the respective populations.

(C) Number of cells recovered by flow cytometry and normalized per embryo for the NKX2-1^{GFP+} populations (lung, thyroid and forebrain) and foregut endoderm. Number of sorts: N=7 for lung, thyroid and forebrain; N=5 for foregut endoderm.
(D) *Nkx2-1* expression (RNA-Seq normalized counts) in sorted Nkx2-1^{GFP+} and Nkx2-1^{GFP}-negative populations.
(E) *Nkx2-1* counts for all *Nkx2-1*$^{GFP+}$ samples (triplicate lung, thyroid and forebrain samples) mapped on the *Nkx2-1* locus. Normalized *Nkx2*-1 counts for each sample are indicated on the right.
(F) Principal Component Analysis (PCA) plot of the eight populations depicted in (B). Arrows connect specified endodermal or ectodermal populations and their respective precursor stages. The partial overlap of the Nkx2-1^{-} lung and forebrain field populations is most probably due to the fact that both populations are heterogeneous and contains mesenchymal, endothelial, neuronal, non-lung foregut, and other lineages.

4.3.1.2. Understanding the genetic program through pathway analysis

The simultaneous purification and characterization of other Nkx2-1^{+} embryonic progenitors and E8.25 precursor populations (foregut endoderm and ectoderm) allowed us to define the genetic program of primordial lung progenitors and signaling pathways that appear to be specific to lung specification, relative to early endoderm and thyroid epithelium (Fig. 14). Previous work in *Xenopus* and mouse has elucidated signals, such as RA, that impart lung competence to foregut endoderm[267,268] as well as lung specification signals emanating from the mesenchyme, such as Bmp and Wnt signals[240,241,250,251].

Our pathway analysis does demonstrate activation of Wnt and Tgf-β superfamily pathways in lung primordial progenitors versus foregut endoderm vindicating the use of Wnt and Bmp activators in in vitro lung specification cocktails for both mouse and human PSC-based systems (Fig. 14A). Although the Shh pathway is also upregulated in primordial lung progenitors, our *ex vivo* foregut explant experiments demonstrate that it acts upstream of β-catenin-dependent Wnt signaling and it may be redundant for *in vitro* PSC-derived foregut specification as long as Wnt activators are present. Interestingly, the retinol metabolism is downregulated in lung progenitors which may imply that RA signaling is necessary for making foregut cells competent to adopt lung fate but it is not required during lung specification as reported previously by Zorn and colleagues[267]. As far as pathways regulating thyroid *versus* lung endodermal fates are concerned, the PI3K-mediated signaling features prominently in thyroid-related (3/10 pathways) but is absent from lung specification-related pathways (Fig. 14A). This is experimentally supported by the inability of FGF ligands to rescue lung ablation following inhibition of Shh signaling in foregut explants as well as previous work from our group showing FGF signaling is required for thyroid but not for lung specification *in vivo* and *in vitro*[228,251]. More specifically, inhibition of PI3K-dependent FGF signaling in either *Xenopus* or in-mouse ESC-derived anterior foregut at the moment of thyroid specification completely abrogated thyroid fate[228]. Intriguingly, the Hippo pathway also appears to be differentially expressed during lung specification (Fig. 14A) and various YAP/TAZ targets such as *Ankrd1 and Cyr61* are upregulated whereas YAP/TAZ negative regulators are downregulated in lung primordial progenitors relative to pre-specified foregut endoderm (Fig. 14B). Yet, the significance of these findings is, at present, unclear as Shh^{Cre}-driven *Yap* deletion within the foregut endoderm does not lead to lung agenesis[269].

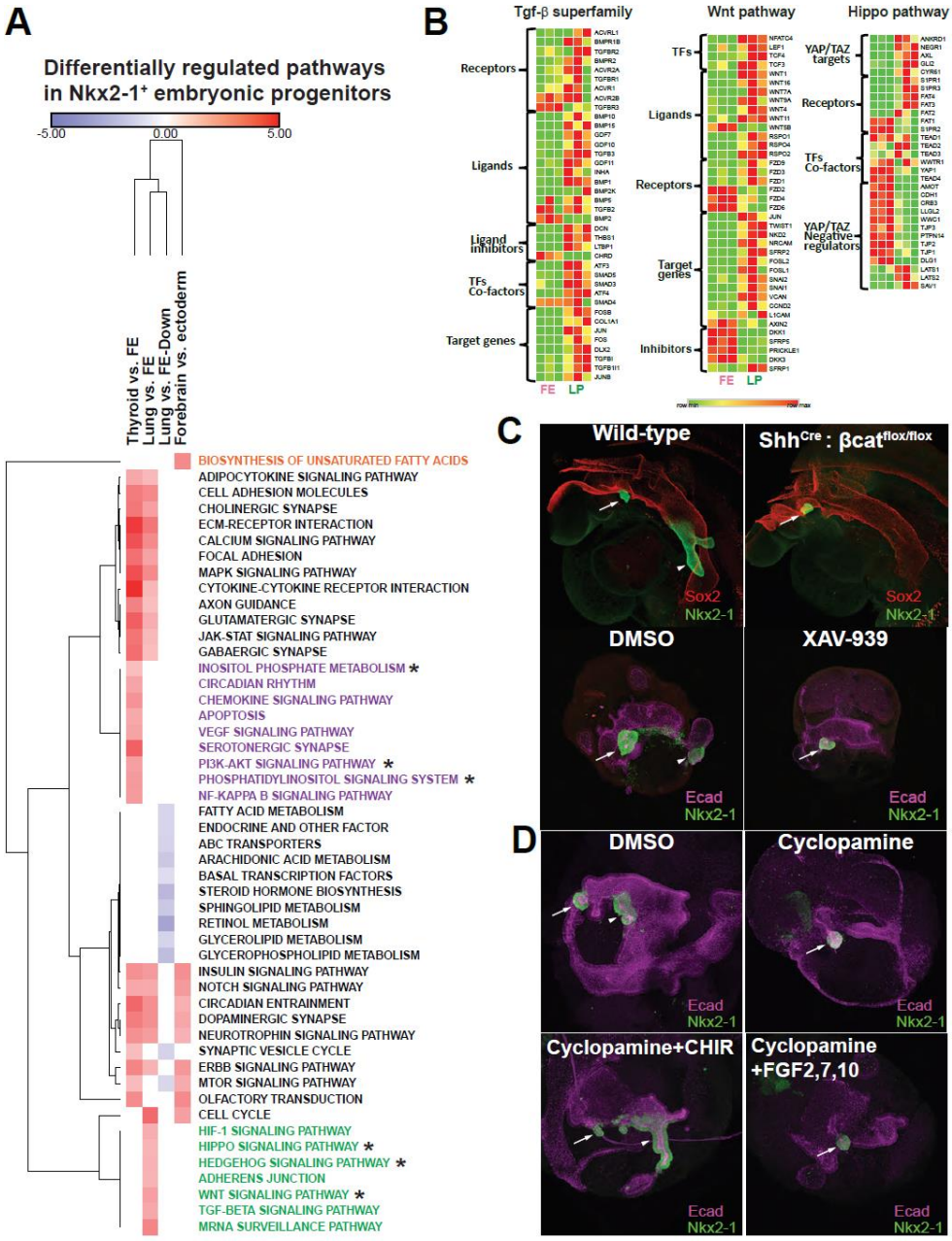

Figure 14: **The embryonic lung specification program,** characterized by differential regulation of the Hippo, Wnt and Tgf-β superfamily pathways.

(A) Heat map of normalized enrichment scores (NES) for pathways that are differentially regulated in specified Nkx2-1+ embryonic populations by gene set enrichment analysis (GSEA). The KEGG pathways database was used as basis for GSEA and an FDR cutoff of 0.25 was used for inclusion of pathways for each pairwise comparison. Pathways that were functionally investigated are indicated with an asterisk (*).

(B) Heat maps of selected transcripts for the Hippo, Wnt and Tgf-β superfamily pathways showing differential expression between pre-specified foregut endoderm and Nkx2-1+ lung primordium.

(C) Whole-mount confocal micrographs of uncultured, freshly isolated embryos at 25-28 somite stage (upper panel) or mouse foregut explants (lower panel). Foreguts were explanted at the 8-somite stage and cultured for 2 days. Nkx2-1 expression demarcates the lung and thyroid domains. Co-staining with Sox2 (upper panel) and E-cad (lower panel) is used to visualize foregut endoderm; Arrow = Thyroid, Arrowhead = Lung.

(D) Whole-mount confocal micrographs of mouse foregut explants. Foreguts were explanted at the 6-8 somite stage and cultured for 3 days. E-cad co-staining marks foregut endoderm; Arrow = Thyroid, Arrowhead = Lung.

4.3.1.3. Comparing *in vivo* and *in vitro* models

Having transcriptionally defined a benchmark for lung primordial progenitors, namely the E9.0 *in vivo* lung epithelial primordial progenitors, this work attempted then to define the fidelity of *in vitro* lung progenitors derived under new biomechanical conditions. As the repertoire of *in vitro* engineered cell types through directed differentiation, direct conversion, and forward programming is rapidly expanding, it is essential to define the correspondence of *in vitro*-derived and *in vivo* cell types. Several methods, including our own work, have been developed that apply network analysis to cell fate conversions[270], identify TFs that underlie cell identity[271,272] and predict reprogramming TFs[238,273]. Previously, we found to what degree to thyroid progenitors derived from PSCs through forward programming were similar to E13.5 thyrocytes[2]. Application of the same methods in the characterization of *in vitro* lung progenitors showed that enhancement of their epithelial properties resulted in increased resemblance to *in vivo* lung primordial progenitors and highly reduced mesenchymal cell signature. Nkx2-1^{+}EPCAM^{+} progenitors derived under 2D and 3D conditions had similar scores that were greatly improved compared to Nkx2-1^{+} progenitors without EPCAM enrichment. This implies that most of the lung competence in the 2D, suboptimal condition appears to be in the small (~1-2%) Nkx2-1^{+}EPCAM^{+} population. Nevertheless, differential similarity scores to epithelial lung progenitors from later time points indicate that culture on different substrata may introduce fate biases, a tantalizing possibility that merits further study. The increased similarity to *in vivo* lung primordium of even the Nkx2-1^{-} EPCAM^{+} population may seem perplexing at first glance, as the absence of Nkx2-1 expression should signify the absence of lung or thyroid endodermal identity. Yet, the fact that this population does not project on other lung epithelial lung populations and has a higher liver signature suggests that Nkx2-1-negative cells are qualitatively different that Nkx2-1^{+} epithelial progenitors at the same stage of *in vitro* directed differentiation. It is also possible that the sorted Nkx2-1^{-}EPCAM^{+} population contains less differentiated cells that were about to adopt a lung fate, as suggested by the higher foregut projection of this population.

In summary, the current work defines the unique genetic program of *in vivo* primordial lung progenitors and uses similarity analysis to increase the correspondence of PSC-derived lung progenitors with their *in vivo* counterparts *via* manipulation of their biomechanical microenvironment. Our findings and the proposed methodology provide a framework for rational and systematic development and refinement of PSC directed differentiation protocols and can propel future studies of lung primordial progenitors.

4.3.2. Sheep's secretome proteomics unravels molecular anti-inflammatory mechanisms

Osteoarthritis (OA), a degenerative joint disease characterized by progressive articular cartilage degeneration, is one of the most commonly diagnosed diseases in general practice and one of the leading causes of disability worldwide[274–277]. In addition to its significant medical, social and psychological impact on quality of life, OA is associated with commensurate socioeconomic costs[278,279]. As adult articular cartilage has little intrinsic repair capacity and current treatment options are mostly palliative, the disease prevalence and burden places a strong emphasis on the need for new therapeutic strategies that could modify the structural progression of the disease and regenerate articular cartilage. The development of disease-modifying anti-OA drugs has thus far proven to be challenging due to the complexity of the disease and the pathophysiological pathways that drive OA progression. While OA has a multifactorial aetiopathogenesis involving genetic, molecular, and biomechanical influences as well as life-style and environmental stress stimuli, it culminates in a consistent molecular, structural and clinical sequence of disease progression, characterized by inflammation, gradual loss of proteoglycans, collagen type II (Col2) degradation, cartilage fibrillation, loss of maturational arrest and phenotypic stability of articular chondrocytes (as reviewed by Goldring *et al*[280] and Pap *et al*[281]).

Fetal mammals, in contrast to adults, are capable of regenerating injured tissues including skin, palate, tendon, bone and cartilage in the first 2 trimesters of gestation[282–290], and fetal scarless regeneration is a paradigm for ideal tissue repair. Despite progress over the past decade, the mechanisms of the tightly regulated process of scarless fetal healing, involving the interplay of growth factors, cytokines, proteinases, and cellular mediators combined with differences in cellular density, proliferation rate, inflammatory response, ECM composition and synthetic function, especially compared to adults remain largely unknown[291–294]. Studies in dermal wound healing identified qualitative, quantitative and temporal differences in growth factor and cytokine expression between adult and fetal wounds [283,295,296].

Experimentation on any humans, and especially fetal subjects, is impossible and unethical. In this study, we aimed to (1) establish a standardized cartilage lesion model allowing comparison of cartilage healing in adult and fetal sheep (*ovis aries*), large animals – physiologically similar to humans in both size and longevity, that should provide a good model for studying human disease; (2) establish the feasibility, repeatability and relevance of proteomic analysis of minute fetal and adult cartilage samples; and (3) compare fetal and adult protein regulation in response to cartilage injury.

The well characterized ontogeny of the ovine immune and inflammatory system and bone marrow niche made the sheep an ideal model in which to examine fetal regeneration[297,298]. The sheep is also a well-accepted and validated model for musculoskeletal disorders and particularly cartilage degeneration[299–303]. Hence sheep are commonly used to study therapies for osteoarthritis

and articular cartilage lesions[304–306]. In addition, fetal sheep share many important physiological and developmental characteristics with humans and have proven themselves invaluable models for mammalian physiology [297,298]. Results obtained in the fetal lamb have been directly applicable to the understanding of human fetal growth and development and are highly predictive of clinical outcome in a variety of applications including in utero stem cell transplantation[297,298,307–311]. Specific characteristics that make sheep particularly well-suited for osteoarthritis, regenerative medicine and fetal regeneration research to obtain results of high clinical relevance are: (1) large size facilitating repeated sampling from individual animals and harvest of adequate sample sizes; (2) comparable bodyweight to humans; (3) similar mechanical exertion to humans[312,313]; (4) relatively long life expectancy (lifespan 8-12 years) allowing longitudinal analysis as well as evaluation of long-term efficacy and safety of treatments; (5) long gestational period (150 days) provides sufficient temporal resolution to translate findings obtained in sheep into human parameters[298]; (6) extremely well characterized immune development analogous to humans[297,298,314–317]; (7) bone marrow ontogeny and niche development closely paralleling humans[298]. Furthermore, for the sheep model, a quite acceptable annotation status and representative subsets of identified proteins are available on sources such as the NCBI (*e.g.*, 30584 genes and 69889 proteins)[318] allowing good applicability and translation of the results.

The proteomic analysis of the differential response of fetal and adult cartilage to injury will have a major impact on our understanding of cartilage biology and of the molecular mechanisms underlying OA and cartilage regeneration, could help identify and target factors that are crucial to promote a regenerative response and may allow the development of disease-modifying treatment strategies to induce cartilage regeneration in adult mammals. A major challenge to the proteomic characterization of the complex protein mixture in cartilage extract is the wide dynamic range of protein abundance, making the detection of low-abundant proteins very difficult[319,320]. However, while technically demanding, studying the functional proteome gives a more comprehensive picture of disease aetiopathogenesis than gene expression analysis alone, as its interpretation is not limited by a possible disconnect between gene and protein expression levels[321].

4.3.2.1. Technical feasibility of the novel sheep model

We first examined and confirmed that the ovine model, in repeatable fashion, supports complex surgical manipulations required for the investigation of cartilage regeneration[1]. All sheep tolerated laparotomy, uterotomy and fetal manipulation well, with no postoperative complications or abortions. With our novel ovine cartilage defect model and analytical approaches we were able to confirm regeneration in fetal versus scarring repair in adult sheep (Fig. 15).

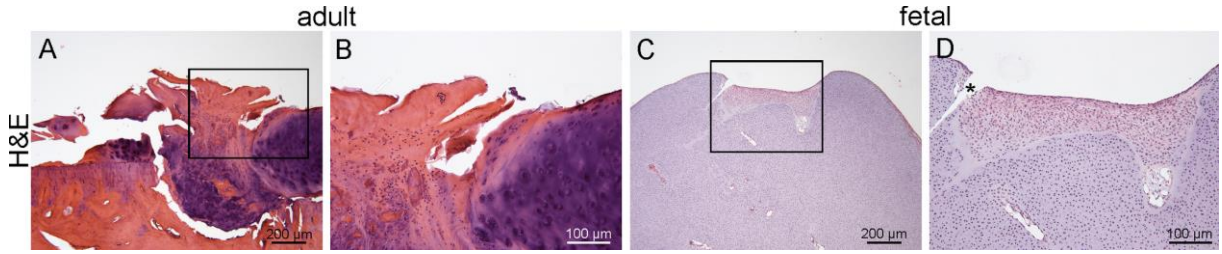

Figure 15: **Healing of adult (5 months post injury) and fetal (28 days post injury) cartilage defects:** Haematoxylin and Eosin (H&E) stained sections. Adult (A and B): no repair except in areas of microfracture, tissue mixture of fibrocartilage with partial hyalinisation, no integration with surrounding cartilage (see insert), Fetal (C and D): defect filled to 80 - 90% with differentiating hyaline cartilage and the superficial 10-20% with repair tissue with progressing hyalinisation, good integration with surrounding cartilage, processing artefact (*).

In this study we compared the adult and fetal response to cartilage injury 3 days after lesion induction as this time point is established to be within the time window of inflammation for both adult and fetal individuals, one of the injury responses hypothesized to crucially contribute to the difference between adult and fetal healing. For the fetal injury response, it is only known that cartilage regeneration occurs within 4 weeks, which is in stark contrast to the adult injury response with an inflammatory phase of 3-5 days, a proliferative phase of 3-6 weeks and a remodeling phase of up to one year duration resulting in a fibrocartilaginous scar. As the timeline of the fetal injury response is not yet established, choosing a later date would have made data interpretation and correlation of adult and fetal data impossible. Three days is within the peak period of the adult inflammatory response, allows for recruitment of monocytes/macrophages to the injury site and has been shown to be associated with signs of inflammation also in fetal injuries in other tissues.

4.3.2.2. Ovine model supports comprehensive molecular profiling via proteomics

Protein identification was performed in an established fashion[1,322] yielding an intensity value per protein. We then developed a custom analysis pipeline, imputing missing values in order to mitigate the effects of random non-observations, normalizing the data by a mean log-shift, standardizing mean expression levels per sample[1]. We then fit linear models separately for each protein, computing second-level contrasts for a direct test of differences between fetal and adult responses to injury[323]. Conservative Benjamini-Yekutieli correction was used to adjust for multiple testing to give strong control of the false discovery rate (FDR). We call significant features for q-values < 0.05. Linear models were adjusted for the nested correlation structure of technical and biological replicates. Significance was assessed by regularized t-tests. For these, group variances are shrunk by an Empirical Bayes procedure to mitigate the high uncertainty of variance estimates for the available sample sizes[324].

Our results demonstrate excellent technical reproducibility, with variation clearly lower than variation between biological replicates, indicating a high sensitivity of the proteomics profiling workflow (Fig. 16). The robustness of our new cartilage defect model is reflected in the variance

across biological replicates being small in relation to the examined biological effects, whether injury versus control, or differences between adult and fetal samples (Fig. 16). For both adult and fetal samples, low variance across replicates indicates good reproducibility of our experimental setup, confirming that biologically meaningful signals can sensitively be obtained already from moderate sample size. Furthermore, it confirms good standardization of our articular cartilage defects between individuals of both the adult and fetal age group.

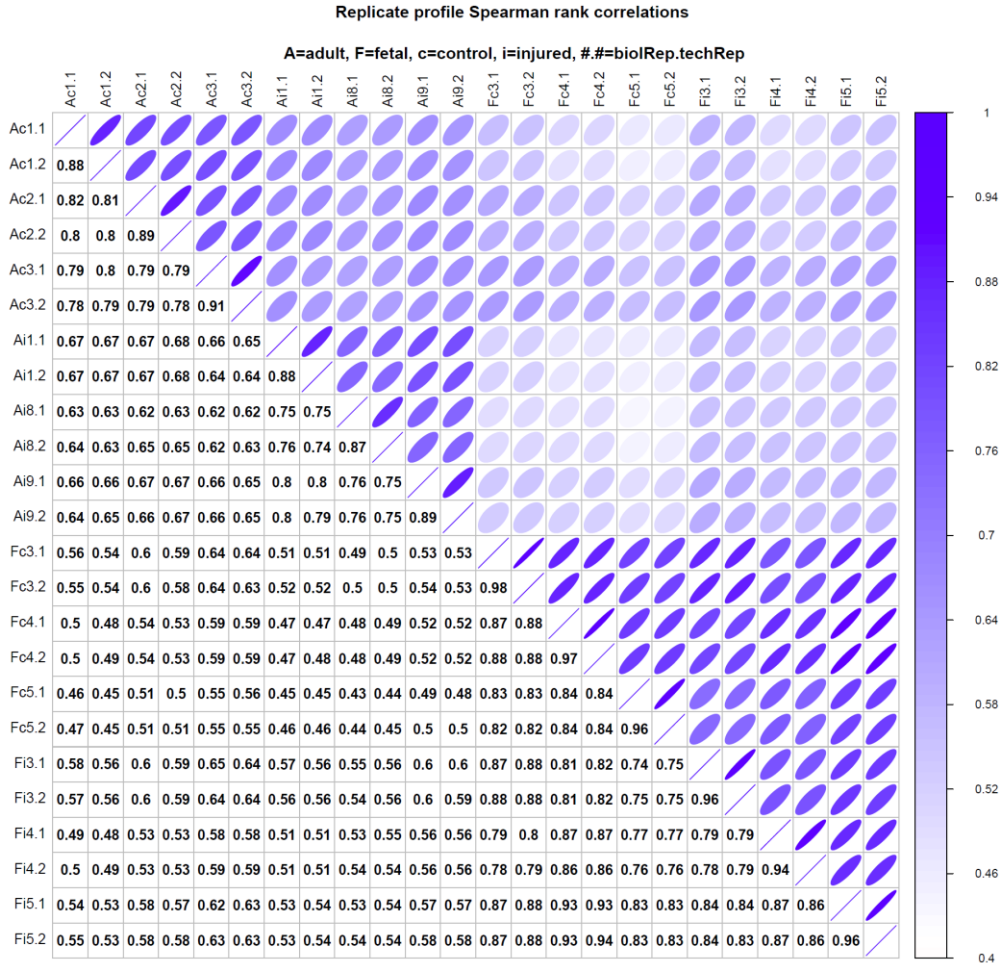

Figure 16: **Sample correlation structure.** This figure compares pairwise sample correlations, with Spearman rank correlation coefficients given in the boxes below the diagonal, which are visualized above the diagonal, with darker and narrower ellipses indicating higher correlations. Rows and columns show sample labels, where A/F=adult/fetal, c/i=control/injured, and #.# show biological and technical replicate numbers (n= 3 biological replicates/group, 2 technical replicates/biological replicate). For the visualization of the sample correlation structure, ellipses are plotted as (x, y) = (cos(θ+d/2), cos(θ-d/2)), where θ ϵ [0,2π) and cos(d)=ρ, with ρ the Spearman rank correlation coefficient[325].

Below we discuss the biological relevance of our new ovine cartilage defect model and MS analysis. Specifically, secretome analysis of control and injured (3 days postoperative) cartilage tissue samples derived from adult and fetal sheep, respectively, using high-resolution mass spectrometry (MS) enabled the identification of a total number of 2106 distinct proteins. Thereof,

445 proteins were found significantly regulated (q-value < 0.05) in response to cartilage injury in adult animals, in contrast to 74 proteins in fetal animals (Fig. 17). Comparing protein baseline expression, 1288 proteins were found significantly differentially regulated between fetal and adult control animals. The injury response of fetal and adult sheep was significantly differently regulated in 356 proteins.

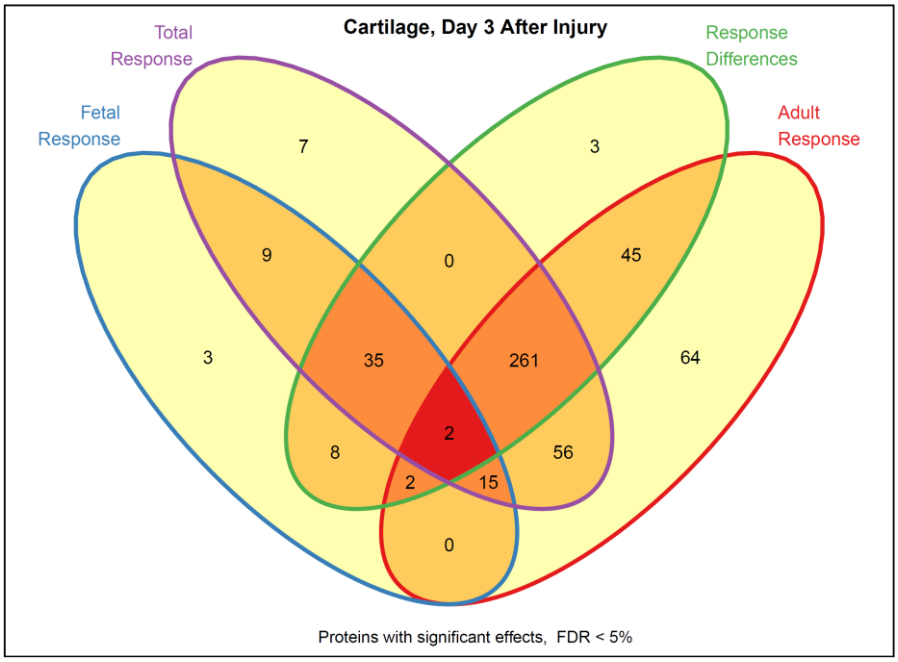

Figure 17: **Proteins implicated by a range of differential screening tests**. Venn diagram gives an overview of proteins implicated by a range of differential screening tests (n= 3 biological replicates/group, 2 technical replicates/biological replicate). Specifically, we examine the average Total Response to injury (magenta), the Fetal Response (blue), the Adult Response (red), and significant Response Differences (green). Separately assessing significance of each of the four tests improves sensitivity and specificity by avoiding an accumulation of thresholding artefacts. Comparing cartilage on day 3 after injury with matching control tissues yielded 385 genes implicated in the total response incorporating the average evidence from all sample types (7+9+0+35+261+2+56+15). Analogously, 74 genes were implicated in the fetal response (9+3+35+8+2+2+15), of which 13 were newly identified (3+8+2+0). Conversely, 445 genes were implicated in the adult response (45+261+64+2+56+2+15), of which 111 were newly identified (45+64+2+0). Response differences are shown by 356 genes with an injury response in fetal samples that was significantly different to that in adult samples (3+0+45+35+261+2+8+2), including 3 previously not implicated genes, 8 genes so far only implicated in the fetal response, 45 genes so far only implicated in the adult response, 2 genes already implicated in both, 35 genes already implicated in the total and the fetal responses, 261 genes already implicated in the total and the adult responses, and 2 genes implicated in all, reflecting that response strength and direction of expression change can be affected differently in the injury response of fetal and adult samples.

The main factors identified within the secretome were extracellular matrix proteins, growth factors and inflammatory mediators such as cytokines and chemokines (Tab. 8). Considering the key chondrocyte signalling pathways regulating processes of inflammation, cell proliferation, differentiation and matrix remodelling, which include the p38, Jnk and Erk Map kinases, the PI-3 kinase-Akt pathway, the Jak-Stat pathway, Rho GTPases and Wnt-β-catenin and Smad

pathways[326], our data provide an initial indication of differences in the inflammatory response to injury between adult and fetal cartilage and suggest the active production of anti-inflammatory and growth factors, such as Ppm1A and Cdc42 in the fetal environment.

Table 8: **Selected relevant proteins.** logFC represents the fold change in a logarithmic scale to the basis 2 based on label-free quantification (LFQ) intensities.

	Fetal ctrl vs Adult ctrl		Fetal D3 inj. vs ctrl		Adult D3 inj. vs ctrl		Fetal vs adult inj. response	
Name	logFC	q	logFC	q	logFC	q	logFC	q
Acan	-10.74	5.54 E-11	1.77	1	-1.36	9.15 E-01	3.13	1.01 E-01
Ccdc88A	4.32	5.29 E-01	-10.45	5.62 E-03	10.05	8.04 E-04	-20.5	1.90 E-05
Ccdc42	-3.58	5.97 E-01	8.68	4.27 E-02	-3.45	9.39 E-01	12.13	5.47 E-03
Chad	-10.94	2.14 E-09	1.48	1	-1.22	1	2.7	6.95 E-01
Col2a1	-2.07	3.69 E-01	1.14	1	-1.01	1	2.15	1
Comp	-9.64	1.13 E-10	1.67	1	-0.05	1	1.72	1
Gabpa	-3.01	3.50 E-02	8.23	1.52 E-05	-0.29	1	8.52	1.09 E-04
Mapk3	-7.81	1.50 E-02	1.11	1.00 E+00	-13.73	1.23 E-05	14.84	2.55 E-03
Ppm1A	-1.62	1	8.91	1.36 E-03	-0.21	1	9.12	4.69 E-03
Prg4	-11.56	3.94 E-12	3.18	1.27 E-02	-1.43	5.29 E-01	4.61	1.63 E-03
S100A12	-2.71	1.86 E-01	8.39	4.99 E-05	13.54	7.37 E-10	-5.15	7.15 E-02
S100A8	-2.78	3.48 E-01	7.49	1.29 E-03	15.8	7.15 E-10	-8.32	3.53 E-03
S100A9	0.08	1	6.34	4.25 E-01	15.45	9.32 E-08	-9.11	4.15 E-02

Ppm1A is a bona fide phosphatase, which is involved in the regulation of many developmental processes such as skeletal and cardiovascular development. Through its role as phosphatase of many signalling molecules such as p38 kinase, Cdk2, phosphatidylinositol 3-kinase (PI3K), Axin and Smad, up-regulation of Ppm1A abolishes for example TGF-β-induced antiproliferative and transcriptional responses[327] as well as BMP signalling[328]. Furthermore Ppm1A by dephosphorylating IκB kinase-β and thus terminating TNFα-induced NF-κB activation, partakes in the regulation of inflammation, immune-response and apoptosis[324].

Cdc42 belongs to the family of Rho GTPases and controls a broad variety of signal transduction pathways regulating cell migration, polarization, adhesion proliferation, differentiation, and apoptosis in a variety of cell types[324]. Cdc42 is required in successive steps of chondrogenesis by promoting mesenchymal condensation via the BMP2/Cdc42/Pak/p38/Smad signalling cascade and chondrogenic differentiation via the TGF-β/Cdc42/Pak/Akt/Sox9 signalling pathway[329]. Another essential Cdc42 function relevant to the current study is its involvement in wound healing by attenuating MMP1 expression[330] and regulating spatially distinct aspects of the cytoskeleton machinery, especially actin mobilization toward the wound[331] which, given the increase of actin-containing articular chondrocytes in response to cartilage injury, could also play a role in the healing of cartilage defects[290].

In contrast to the anti-inflammatory factors up-regulated in fetal sheep in response to injury, adult sheep displayed a significant increase of inflammatory mediators such as alarmins S100A8, S100A9, S100A12 and coiled-coil domain containing 88A (Ccdc88A) (Tab. 8). The alarmin S100 proteins are markers of destructive processes such as those occurring in OA[332–334]. Accordingly, in OA articular S100A8 and S100A9 protein secretion is increased, recruiting

immune cells to inflamed synovia, initiating the adaptive immune response, inducing canonical Wnt signalling and promoting cartilage matrix catabolism, osteophyte formation, angiogenesis and hypertrophic differentiation[332–334]. S100A8/A9 up-regulate markers characteristic for ECM degradation (MMPs 1, 3, 9, and 13, interleukin-6 (IL-6), IL-8) and down-regulate growth promotion markers (aggrecan and Col2) and thus have a destructive effect on chondrocytes, causing proteoglycan depletion and cartilage breakdown[335]. Also S100A12 is up-regulated in OA cartilage and has been shown to increase the production of MMP-13 and Vegf in OA chondrocytes via p38 Mapk and NF-κB pathways[336]. Another relevant protein, which was significantly down-regulated upon injury in fetal sheep but significantly up-regulated in injured adult sheep is Ccdc88A. Ccdc88A is a multimodular signal transducer, which modulates growth factor signalling during diverse biological and disease processes including cell migration, chemotaxis, development, self-renewal, apoptosis and autophagy by integrating signals downstream of a variety of growth factors, such as Efg, Igf, Vegf, Insulin, Stat3, Pdgfr and trimeric G protein Gi[337,338]. In addition, Ccdc88A, which is expressed at high level in immune cells of the lymphoid lineage, plays an important role in T cell maturation, activation and cytokine production during pathological inflammation and its inhibition could help treat inflammatory conditions as shown in in-vitro and mouse studies[339]. Furthermore Ccdc88A, via activation of Gαi, simultaneously enhances the profibrotic (Pi3k-Akt-FoxO1 and TGF-β-Smad) and inhibits the antifibrotic (cAMP-PKA-pCREB) pathways, shifting the fibrogenic signalling network toward a profibrotic programme[340]. Interestingly, in the liver, sustained up-regulation of Ccdc88A occurs only in all forms of chronic fibrogenic injuries but not in acute injuries that heal without fibrosis, indicating that increased expression of Ccdc88A during acute injuries may enhance progression to chronicity and fibrosis[340]. Ccdc88A also regulates the Pi3 kinase-Akt pathway, which exhibits pleiotropic functions in chondrogenesis, cartilage homeostasis and inflammation. It may further induce an increase in MMP production by chondrocytes leading to subsequent cartilage degeneration, via its multiple downstream target proteins[341–347].

Remarkably, in this study, the cartilage matrix proteins Prg4, Acan, Comp and Chad had a significantly higher baseline expression in adult sheep and showed little injury response in either age group with the exception of Prg4, which was significantly up-regulated in fetal injured sheep (Tab. 8). Prg4, in response to injury increased 3.2 fold (q=0.01) in fetal sheep, which is a 4.6 fold higher increase compared to adults (q=0. 002), indicating a stronger and more rapid cartilage matrix production. Since Prg4 expressing cells constitute a cartilage progenitor cell population, the higher baseline expression in adults is particularly surprising but can be explained by its restriction to the most superficial cell layer in fetal joints compared to a distribution throughout the entire cartilage depth in older individuals[348].

In contrast to the cartilage matrix glycoproteins, many growth factors, such as Gabpa and Mapk3 showed, as expected, differential regulation following injury between adult and fetal sheep (Tab. 8). Gabpa, a member of the ets protein family, which is ubiquitously expressed and plays an essential role in cellular functions such as cell cycle regulation, cellular growth, apoptosis, and

differentiation[349] showed a further significant up-regulation in fetal injury and no response to adult injury (q=0.0001). Gabpa activates the transcriptional co-activator Yes-associated protein (Yap), which is essential for cellular and tissue defences against oxidative stress, cell survival and proliferation and can induce the expression of growth-promoting genes important for tissue regeneration after injury[350]. The cellular importance of Gabpa is further highlighted by the observation that in Gabpa conditional knockout embryonic stem cells (ESCs), disruption of Gabpa drastically repressed ESC proliferation and cells started to die within 2 days[351].

The growth-regulator Mapk3 had a higher baseline expression in adult sheep (logFC=7.8, q=0.02) but significantly decreased (13.7 logFC, q<0.0001) after injury, while fetal Mapk3 remained essentially unchanged (Tab. 8). Mapk3 acts as an essential component of the MAP kinase signal transduction pathway and as such contributes to cell growth, adhesion, survival and differentiation through the regulation of transcription, translation and cytoskeletal rearrangements. Mapk3 also fulfils an essential role in the control of chondrogenesis and osteogenesis of MSCs under TGF-β or mechanical induction and positively regulates chondrogenesis of MSCs[352].

4.3.2.3. Sheep as a vehicle to understanding selected human diseases

Our results are consistent with previous studies in fetal skin wounds, which also have shown a different and reduced inflammatory response and decreased scar formation during dermal wound healing in fetal sheep and mice[295]. The chronic progressive articular damage of OA is associated with similar levels of pro-inflammatory cytokines and chemokines throughout the disease[321] and occurs via a complex program involving on-going local inflammation triggered by cytokines and endogenous activation of innate immunity, complement and metabolic pathways[281]. Therefore, the different inflammatory response as we have demonstrated in the fetus may be a major contributor to fetal scarless cartilage healing. This is especially intriguing as fetal sheep have a normally functioning immune system by 75 gd[297,353]. Leukocytes have been shown to be present and increase rapidly at the end of the first trimester[314,354]. Fetal sheep are able to form large amounts of specific antibodies in response to antigen stimuli by 70 gd[355] and reject orthotopic skin grafts and stem cell xenotransplants administered after 75-77 gd with the same competence and rapidity as adult[356]. Furthermore, fetal sheep have an inflammatory response to injury before 80 gd[357–359]. The first evidence of inflammation, the presence of TNF and IL-1 has even been shown as early as 30 - 40 gd[360].

In conclusion, our results demonstrated the power of the sheep model organism to facilitate understanding a corresponding disease in humans. Specifically, with our novel ovine fetal cartilage regeneration model and analytical approach, both positive and negative regulators of inflammatory events were found to be differentially regulated. This holds promise for potential therapeutic interventions as the presence of a negative regulator is more easily mimicked than the absence of a positive regulator. Further studies employing this newly developed animal model

and analytical techniques to identify proteins involved in OA aetiology and pathogenesis, as well as potential biomarkers and therapeutic targets are warranted to work toward the goal of novel biomimetic solutions, which might be exploited to favourably shift the adult cartilage healing milieu to a more fetal phenotype to induce regeneration by recapitulating cartilage ontogeny.

4.4. Patient samples reflecting variation from non-uniform cohort structure

As we have shown, both cell lines and model organisms are well suited to small scale but highly dimensional experiments, improving the state of knowledge in areas, like, targeted cell differentiation[2], or tissue regeneration, potentially transferable to inflamed human tissue[1]. However, with these highly dimensional quantitative measurements, employed to characterise the heterogeneous *in vivo* models, and increasing biological variation, the small sample size becomes an issue[361]. Especially, in the treatment of cancers, predicting clinical outcome and selecting an effective personalized therapy for individual patients is particularly challenging due to the intrinsic heterogeneity of the disease[362,363]. For instance, the most commonly occurring malignancy in women – breast cancer, is known to include multiple distinct molecular subtypes with disease progression and treatment response varying widely across patients[364–366]. Because of this complexity of cancer patient cohorts only larger sample size can enable us to try to explain this disease better and facilitate finding optimized, or personalized, treatments[363].

4.4.1. The role of data integration in overcoming limitations from unwanted variation

Furthermore, integration of measurements generated with multiple varying technologies should improve sensitivity to small changes in signal[29], and help finding different aspects of biology from complementary information. Combining different assay types may capture complementary aspects of information and, investigated across patients, could help shed new light on complex relationships of interest, such as the relationship between genotype and eventual consequences for cancer progression. Importantly, because each genome-scale data type has different characteristics, with different type-specific random variation, or noise, present, and bias introduced by multiple different factors[94,166,168–177], data integration should also facilitate overcoming these issues inherent in individual studies[30]. Notably, different measurement characteristics of each molecular profile type need to be taken into consideration when analysing data jointly[47]. This is important, because by introducing the additional measurement noise from a different data source data integration could even be detrimental. Specifically, the additional noise from a new data source can dominate overall results when other data sources contribute considerably more information compared to the information added by the new data source[48]. A successful robust integrated data analysis pipeline therefore needs to adapt to situations where an integration of data may be beneficial *versus* situations where additional noise from an additional

data source just deteriorates the overall signal, a challenge when improving and developing novel integrative methods.

In the following chapter, I show that complementary measurements may facilitate clinically relevant patient stratification or accurate survival time prediction.

5. Chapter 5: Integration of heterogeneous data sources

In this chapter, I first analyse cancer cohorts available in a large neuroblastoma study[367–371] and the TCGA database[28]. I focus on cancer because of the wide availability of matched multi-modal data for this group of diseases. TCGA itself encompasses over thirty cancers, hosting multiple independent studies per cancer, with multiple matched molecular profiles, like RNA-Seq, copy number, or methylation data, and corresponding clinical information. I exploit the data for stratifying patients into clinically-relevant treatment groups, and predicting survival time, applications relevant to improvement and personalization of the treatment of cancer patients. I always perform single-track analyses, involving a specific molecular profile, to then compare these results with results obtained with my newly developed integrative algorithms. Importantly, I establish a novel benchmarking metric and compare the performance of my models with state-of-the-art algorithms and independent results presented in topical literature.

5.1. A multi-layer network approach to data integration for patient stratification

5.1.1. Patient classification

In the treatment of cancers, selecting an effective therapy for individual patients can be particularly challenging because of the heterogeneity of the disease, where disease progression and treatment response can vary widely across patients. In order to predict a patient's response clinicians delineate subgroups of patients likely to react similarly. Typical criteria of classification include clinical records, such as the age at diagnosis, sex, and comorbidities. It is by now widely recognized that the underlying mechanisms can vary widely across patients. Even for patients with the same specific clinical diagnosis, for instance, we find many subtypes of breast cancer [16] or adult acute myeloid leukemia (AML)[17]. Combining clinical records and histologic tests alone often cannot reliably identify the biological processes underlying a particular tumor type[18]. Increasingly, therefore, molecular markers are now incorporated to improve the prediction of therapy response and prognosis[18–21]. Common molecular markers include changes in gene activity, such as identifying characteristic gene sets or signatures[22,23], and genomic sequence variants, such as copy number changes from deletions and amplifications of certain genomic

regions, like the amplification of the MYCN gene in Neuroblastoma patients[24], as well as smaller changes, such as single nucleotide polymorphisms[25].

5.1.1.1. Improvement through data integration

With the very large amounts of molecular data now collected from biomedical assays, drawing clinically relevant insight has soon become the bottleneck that still remains rate-limiting. A lot of hope is being placed in analysis approaches which combine measurements from different sources, be that horizontally, across patients[29,31], or vertically, across assay types[31–33]. Vertical integration combines different assay types that may capture complementary aspects of information and, investigated together, could help shed new light on complex relationships of interest, such as the relationship between genotype and eventual consequences for tumour progression. Specifically, one would expect the impact of gains and losses of gene copies to affect the expression of cis and trans genes of relevance to cancer in a non-trivial way, and complementary measurements may facilitate accurate survival time prediction.

Employing both gene expression and copy number information was reported to improve clustering for subtype analysis using a probabilistic model of joint latent variables[36,37]. Introducing dimensionality reduction by low-rank approximation, LRAcluster was reported to further improve the stratification of cancer patients[42], establishing a probabilistic model of different data types conditional on shared latent factors.

5.1.1.2. Network-based approaches

Alternative approaches include network-based algorithms, which have seen a variety of successful applications, such as the identification of dysregulated pathways[43,44] or an optimization of biotechnological processes[45]. In the Similarity Network Fusion (SNF) algorithm, network representations exploit not just the complementary nature of data sources but also similarities across patients[46], *i.e.* combining both vertical and horizontal data integration. First, patients are grouped into a network, with every node representing a patient. Distances in the network reflect the similarity of the patients in respect of a particular data type. These networks for the different data types are then merged through cross-diffusion by message passing across the networks. This effectively combines evidence across complementary data types and patients, with the hope of identifying relevant patterns of the underlying biology of the disease.

We here introduce a novel network-based approach for the integration of multiple molecular data types. By design, our algorithm extracts information at multiple levels, starting with differential effect analysis of the molecular data. Going beyond earlier work, our approach then directly incorporates *functional knowledge* from expert curated sources, including GeneOntology (GO)[213,214], which lists genes involved in known biological processes, as well as pathway annotations from the KEGG[215] database. Finally, similarities across patients are exploited to build a multiple-layer network[372] that captures all the structured information discovered. This

consolidated resource can then be exploited at various levels, be that the exploration of functional modules identified across patients, or the identification of clinically relevant groupings of patients in the network structure.

5.1.1.3. Challenges with performance assessment

In the context of predicting patient survival, commonly a group of similar patients is identified that can be associated with an average risk change. Thus a grouping of patients is clinically relevant if one can find a sufficiently reduced or increased risk for a reasonable number of patients with significance. Notably, significance alone is insufficient for clinical relevance[52]. Although there is no universally agreed threshold for the clinical relevance of hazard ratios for death by cancer, a risk change of 14% is typically considered to be small, changes of 47-90% can be considered moderate size effects, while risk changes of 90% or more are considered large[53]. Large effects can often be identified more easily for more specific, smaller subsets of patients. For a meaningful comparison of different predictions we therefore need to consider the number of patients as well as the size of the risk change. To this end we introduce an *effective number of affected patients* and demonstrate its use as a balanced metric.

We explore the value of our approach on a range of cancers and data types using several complementary assays.

5.1.2. A novel algorithm for the construction of patient similarity graphs

We here introduce a new algorithm relating patients based on matched molecular profiles, such as gene expression, copy number information, or others. Rather than considering patient similarity based on the molecular profiles alone[46], we combine these with knowledge of biological processes (GeneOntology, GO[213,214]) and different classes of pathways (KEGG[215]) into a multi-level patient / pathway network. Together, this allows for the first time a systematic exploration of relationships incorporating both functional knowledge and multiple molecular profiles.

First, pathway networks are built for each patient and type of molecular profile separately (Fig. 18A), directly incorporating functional knowledge from GO and KEGG[43]. In comparison to methods based only on molecular profiles, the newly introduced functional focus reduces dimensionality to extract higher-level actionable patterns, and attenuates noise. Specifically, at this point, each patient is represented by a network of KEGG pathways that are connected by weighted edges. Weights are computed from the probabilities of the pathway-associated genes (KEGG) showing profiling differences between the tested patient and controls (*e.g.*, differential expression), while considering co-occurrence of genes in the same molecular processes (GO). An integrated view of KEGG pathways can be obtained by summation over GO processes. A stochastic block model (SBM)[373,374] is then used to find robust[375] pathway clusters in that network. This set of pathway clusters can be used to characterize *the disease of a specific patient* as reflected in a particular molecular profile type and already incorporating functional

knowledge. The shared information distance between the clusterings[376] for different patients is then a natural metric of patient similarities. The shared information distance of clusterings is also known as the Variation of Information (VI) metric, and has favourable robustness and locality properties [377]. These are distances corresponding to similarity between two nodes in the network, *i.e.* how closely related two patients are with each other. The smaller the VI and shorter the distance, the stronger the similarity between two nodes. In order to obtain a weight corresponding to this distance we further subtract the VI from 1. This way we arrive at the *weight*-based integrated patient similarity graph, where the higher the weight between two nodes the stronger the similarity between the two patients.

The metric furthermore allows a meaningful direct comparison and, thus, combination of distances. We exploit the first property to construct a robust patient similarity graph (Fig. 18B) for each molecular profile type (gene expression, copy number, …). In summary, we obtain vector-valued edge weights between patients with each vector coordinate corresponding to a particular molecular profile type. We then take advantage of the second property to combine information across different molecular profile types, and average the VI values, yielding a single *integrated patient similarity graph* (Fig. 18C). After the VI distances are integrated, we normalize them as suggested by Meila[377]. Specifically, we multiply each VI value by $1/\log N$, where N is the number of data points in the specific cancer data set, *i.e.* cohort size or number of patients. This bounds the range of values between 0 and 1. As discussed, in order to obtain a weight corresponding to this distance we further subtract the normalized VI from 1.

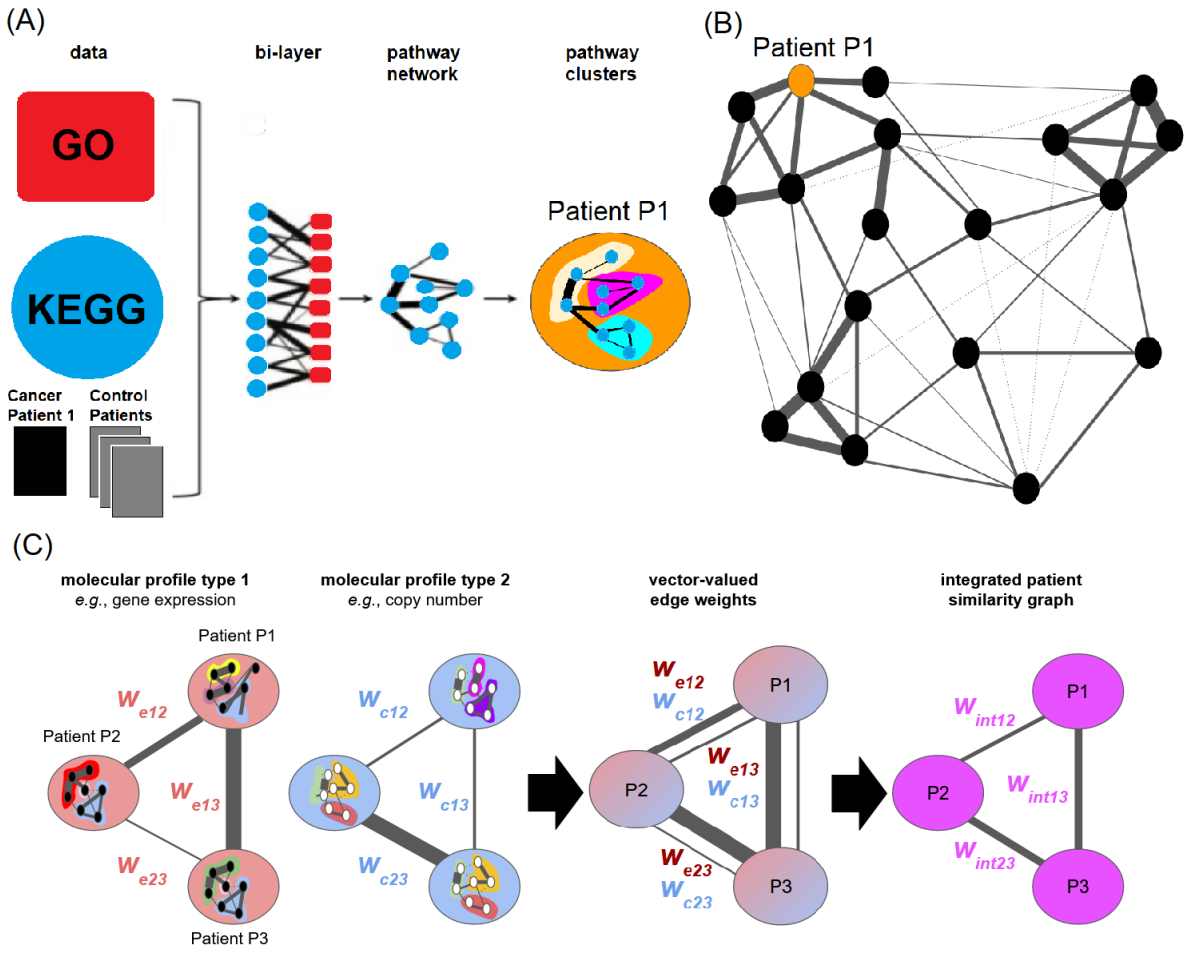

Figure 18: **Overview of the network structures exploited by the novel algorithm.**

(A) **Building a patient-specific pathway network from a single molecular profile type using functional knowledge.** The patient-specific disease profile (black) is compared to one or several controls (grey). The resulting differential effect (*e.g.*, differential expression) is then assessed for genes of known biological processes (GO, red) in a range of KEGG pathways (blue). The magnitude of the posterior probabilities of a differential effect for these genes then determines the weights for a bi-layer network of KEGG pathway and GO biological process nodes. The weights show the magnitude of the posterior probabilities that the shared genes exhibit a differential effect (*e.g.*, differential expression). After integrating the bi-layer network by summation, each patient is represented by a network of KEGG pathways connected by weighted edges that reflect the similarity between pathways in the context of the patient's disease. The robust pathway clusters identified by a stochastic block model (shown in three colours for Patient P1) can be used as a characteristic fingerprint. Colour indicates node type: red=GO, blue=KEGG. Weights are illustrated by black lines with thickness representing weight strength.

(B) **Building a patient similarity graph from a single molecular profile type using pathway clusters.** The set of pathway clusters for a patient can be used to characterize the disease of that patient as reflected in a particular molecular profile type and already incorporating functional knowledge (Fig. 18A). The shared information distance between the clusterings for different patients then forms a natural measure of patient similarities, the Variation of Information metric. Constructing a network of patients based on this metric thus allows an exploration of similarities between patients in the context of the disease. Nodes in the *patient similarity graph* represent patients and edge weights reflect pairwise similarities. The figure omits some connections for visual clarity. The thicker a connecting edge, the stronger the similarity.

(C) **Integrating multiple molecular profile types.** A patient similarity graph is first constructed for each molecular profile type (*cf.* Fig. 18A). Nodes in the graph represent patients and edge weights reflect the patient similarities for

each molecular profile type. The different graphs can be superimposed to give one graph with vector-valued weights for edges between patient nodes. An integrated patient similarity graph is then obtained by combining the information across molecular profile types for edges reflecting an average patient similarity (see Methods).

This integrated patient similarity graph now captures all the structured information discovered both from individual patients and across patient cohorts, as guided by the functional focus of incorporating knowledge about pathways and biological processes. The structure resulting from this *Variation of information fused Layers of Networks* (ViLoN) can then be exploited at various levels, be that in an exploration of functional modules identified across patients, the establishment of more powerful prognoses, or information rich patient stratification for precision medicine, seeking an effective assignment of patient specific treatments. We here demonstrate the effectiveness of clustering patients in the integrated patient similarity graph using a Stochastic Block Model (SBM) to successfully identify similar patients with shared risk profiles. Survival analysis confirms that we can find patient groups of high clinical relevance, where many patients are affected by high hazard ratios.

5.1.3. A metric for clinically relevant patient stratification

Patient stratification seeks to identify a group of similar patients that can be associated with an average risk change. Such a grouping of patients is clinically relevant if the reduced or increased risk is sufficiently large, affects a reasonable number of patients, and passes statistical significance tests. While significance alone is insufficient for clinical relevance[52], there is no universally agreed threshold for the clinical relevance of hazard ratios for death by cancer[53]. Large effects can often be identified more easily for more specific, smaller subsets of patients (cf. Fig. 19, right panel). For small groups, we could even observe hazard ratios above 10^{10} (data not shown). Conversely, small hazard ratios may be statistically significant but clinically not actionable. For a meaningful comparison of different splits of a patient cohort into groups we therefore need to consider the number of patients as well as the size of the risk change. To this end we introduce an *effective number of affected patients* as a balanced metric. Specifically, the product of the risk change and the number of affected patients gives a score reflecting an effective number of affected patients. To also allow a comparison of studies restricted to a subset of the full cohort (or other cohorts of different sizes), we standardize the score to a cohort of 1000 patients,

$$N_{eff} = 1000 \, \frac{N_g \left(2^{\,\mathrm{abs}\left(\log_2 HR\right)} - 1 \right)}{N} \qquad (1)$$

for a hazard ratio *HR* and the affected group size *Ng* in a cohort size *N*.

Hazard ratios are always obtained for a specific subset of patients compared to a distinct reference group of patients providing a baseline. It then is the size of the smaller of the two

groups that determines how meaningful an observed high hazard ratio actually is. For a fair comparison of different methods and their results, for each patient clustering we make the conservative choice of selecting the largest group as the reference baseline, avoiding an artificial inflation of hazard ratios.

Notably, we perform all the survival analyses applying Cox regression with Firth correction for right-censored data that has proven to be reliable, opposed to normal Cox regression, when information is largely censored[378].

5.1.4. Performance in the CAMDA cancer data integration challenge

The annual CAMDA data analysis challenge (www.camda.info) provides a well recognized forum for open-ended comparative exploration of novel algorithms[379–381]. We here consider the neuroblastoma dataset[367,368,370,371]. Neuroblastoma is the most common cancer in children (www.cancer.gov/types/neuroblastoma/), causing 15% of cancer-related deaths at a young age[382]. Despite modern therapies less than 40% of high-risk patients survive[383]. While some patients show spontaneous recovery it remains hard to predict who is at risk and assign appropriate therapy. A more precise prognosis and more effective treatment assignment is now expected to require an integration of molecular patient profiles[18–21].

In this cohort, three matched molecular profile types are available for 145 patients: RNA-Seq, microarray gene expression, and copy number data. Clinical records of both 97 low and 48 high risk patients include the survival times of the patients, with the majority of observations being right-censored, *i.e.* at the end of the study the patient was still alive or dropped off the study before it ended.

5.1.4.1. Robustness of patient stratification

Groups of patients found by ViLoN were stable to removing patients, *i.e.* the network was robust to sample changes. When we remove a patient from the data set, the remaining patients mostly fall into the same clusters as obtained from the complete patient cohort. Leave-one-out robustness test, removing each patient in turn, yielded an average accuracy of 98% (SE 11%) and an average normalized mutual information[384] of 97% (SE 6%). Values of 100% designate perfect agreement.

5.1.4.2. Comparison to previous work

ViLoN results for individual molecular profile types compare favourably with the most effective patient stratifications reported in the literature for gene expression profiles and copy number data[367–371]. Several complementary statistics and metrics of model performances are displayed in Tab. 9. A small *p* value indicates statistical significance and the lower the Akaike Information Criterion, the better the model fit. The latter value is notably not always available for published patient splits. In all cases, however, we can consider the hazard ratio (HR) that characterizes the risk difference between the patient groups, and the number of patients affected, which can be

used to construct an *effective number of affected patients* as a balanced metric N_{eff} (1). The clinically most relevant grouping will have a high hazard ratio and a large number of affected patients, as reflected in a large N_{eff} score.

By all criteria, ViLoN results for individual molecular profile types already show a clear improvement over the best patient stratifications reported earlier, with results integrating information across profile types performing even better.

Table 9: **Comparison to the best reported patient stratifications.** The table shows, method: name of the method for which results are listed in the corresponding table row; data type: type of data used in the analysis; *p*: *p*-value indicating statistical significance; AIC: goodness of the model fit assessed by Akaike Information Criterion (AIC, lower values are better); grSize: size of the patient group with the largest hazard ratio in the model; fullSize: size of the whole analyzed patient group; absHR: the largest absolute hazard ratio; $\log_2$HR: a corresponding $\log_2$ of the hazard ratio - symmetrical for increased/decreased risk; N_{eff}: the effective number of affected patients score.

method	data type	*p*	AIC	grSize	fullSize	absHR	log2HR	N_{eff}
ViLoN	int	<10^-9	317.1	68	145	11.08	3.5	4727
ViLoN	acgh	<10^-7	328.7	69	145	6.59	2.7	2660
Theissen	acgh	-	-	93	202	5.07	2.3	1874
ViLoN	marray	<10^-6	329.0	54	145	5.62	2.5	1721
Kocak	marray	<10^-3	-	208	574	4.44	2.2	1247

By all criteria, representing patient data as a network, combining information across patients already gave excellent performance for individual molecular profile types, improving on results reported in earlier work. Vertically integrating patient networks across molecular profile types considerably further improved stratification performance.

We then compared the performance of ViLoN with alternative stratification algorithms: LRAcluster[42] and SNF[46], and the best model developed for the CAMDA challenge[385]. Even though other integrative approaches yielded excellent results, incorporating functional domain knowledge from KEGG/GO and combining information across patients, exceled these other methods (Tab. 10).

Table 10: **Comparisons of performance of integrative algorithms in 2-group stratification.** ViLoN performance on the Neuroblastoma dataset compared to the largest hazard ratios found by LRAcluster, SNF, and best CAMDA-developed model. The table shows: method: name of the method for which results are listed in the corresponding table row; data type: type of data used in the analysis; p: *p*-value indicating statistical significance; AIC: goodness of the model fit assessed by Akaike Information Criterion (AIC, lower values are better); grSize: size of the patient group with the largest hazard ratio in the model; refSize: size of the largest, *i.e.* reference, patient group in the model; fullSize: size of the whole analyzed patient group; absHR: the largest absolute hazard ratio; $\log_2$HR: a corresponding $\log_2$ of the hazard ratio - symmetrical for increased/decreased risk; N_{eff}: the effective number of affected patients score.

method	data type	p	AIC	grSize	refSize	fullSize	absHR	log_2HR	N_{eff}
ViLoN	int	<10^-9	317.1	68	77	145	11.08	3.5	4727
Kazan	int	<10^-9	NA	172	326	*498*	6.39	2.7	1863
SNF	int	<10^-6	325.1	45	100	145	6.15	2.6	1598
LRAcluster	int	<10^-4	336.4	48	97	145	4.17	2.1	1049

Our new approach not only improved on results reported in earlier work but also performed very well compared to recent alternative approaches to data integration, including the probabilistic dimensionality reduction algorithm (LRAcluster), another established network-based method (SNF), and the best method developed and benchmarked specifically on this dataset for a recent CAMDA competition (Kazan) (Fig. 19A). Interestingly, a method aiming to identify cancer driver genes[386], which may point to new therapy approaches, yielded lower stratification scores (data not shown), suggesting that drivers may not necessarily be the most prominent markers for patient stratification in the context of currently available therapies.

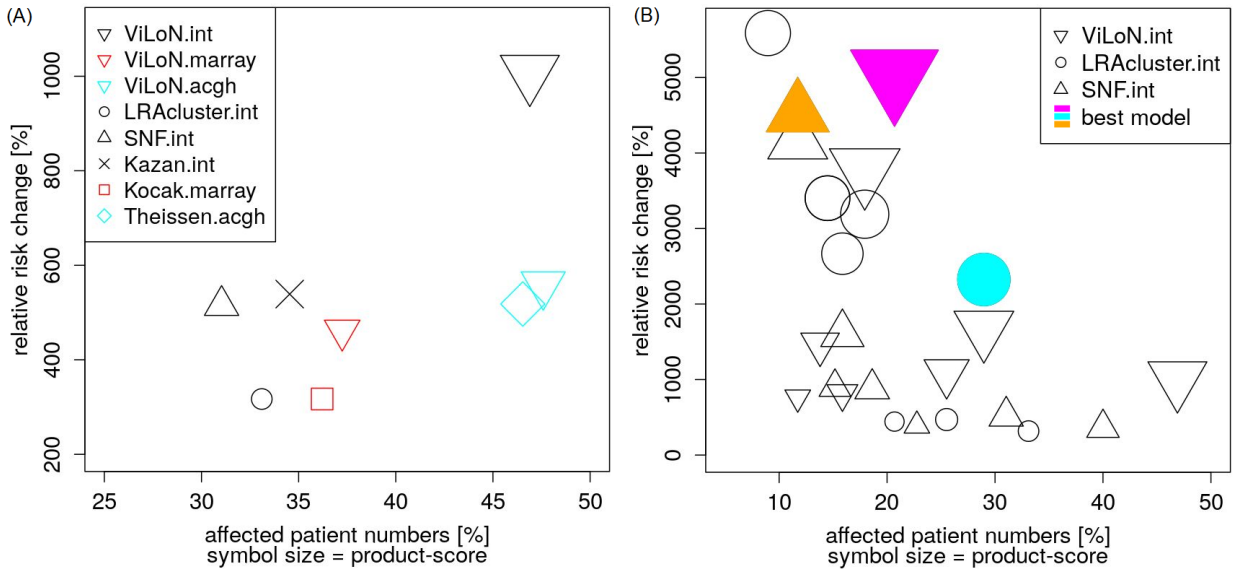

Figure 19: **Comparison of effective number of affected patients scores.**
(A) **2-group stratification.** We compare ViLoN (triangle, point down) with two novel integrative algorithms: LRAcluster (circle) and SNF (triangle, point up); results reported in relevant literature: 'Kocak' (square), 'Theissen' (diamond); and best CAMDA model: Baali (cross). Colors represent different molecular profiles, with black color symbolizing integration. The larger the plotted symbol the larger the product.
(B) **Integrative analysis, with *N*-group stratification.** We compare ViLoN (triangle, point down) with LRAcluster (circle) and SNF (triangle, point up) algorithms. Shown are all the significant models for *N* between 2 and 9, independent of *N*. The larger the plotted symbol the larger the product, with colors accenting the best achieved scores by each method.

The grouping by ViLoN, interestingly, already covers most of patients identified by SNF (43/45 = 96%) while expanding the group of high risk patients considerably (+25 = +58%) (Fig. 20). At the same time, the risk increase associated with group membership is not diluted and the group-associated risk is even considerably sharpened from 6.2 to 11.1 (Tab. 10).

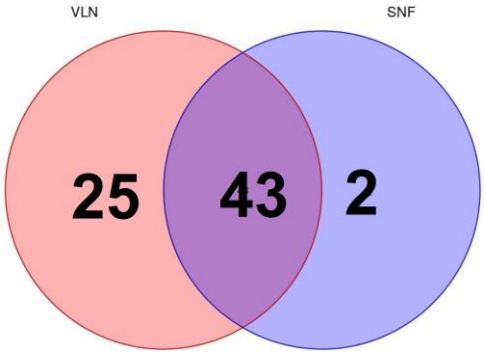

Figure 20: **Comparison of groups found by ViLoN or SNF algorithms**. A Venn diagram compares groups stratified by the alternative algorithms, in a 2-group stratification. Grouping by ViLoN covers most of patients identified by SNF (43/45) and expands the group of high risk patients considerably (+25)

The standard benchmark Cox-Hazard model estimates the hazard ratio between two groups, allowing a direct comparison of a range of stratification methods, also relative to clinical reports in the literature. Several modern methods, however, can stratify into multiple groups, which can independently be assessed relative to a baseline, using the largest group as the reference. Indeed, considering stratification results from 2 to 9 clusters, the best pairwise performances for the compared integrative methods are found for a group of patients identified in a stratification by SNF, LRAcluster, and ViLoN into 8, 3, and 5 clusters, respectively (Tab. 10, Fig. 19B). Note that the model with the largest product score and thus the largest effective number of affected patients, and so clinical relevance, is not necessarily the best fitting model overall.

Table 11: **Best overall performance compared to the state-of-the-art clinical neuroblastoma grouping.** The table shows: method: name of the method for which results are listed in the corresponding table row; data type: type of data used in the analysis; #g: number of strata the model stratified patients into (between 2 and 9); adj.p: *p*-value of the individual model with the largest hazard ratio in the whole model, adjusted for multiple testing, indicating statistical significance; adj.p.model: *p*-value of the whole model adjusted for multiple testing, indicating statistical significance; AIC: goodness of the model fit assessed by Akaike Information Criterion (AIC, lower values are better); grSize: size of the patient group with the largest hazard ratio in the model; refSize: size of the largest, *i.e.* reference, patient group in the model; fullSize: size of the whole analyzed patient group; absHR: the largest absolute hazard ratio; $\log_2$HR: a corresponding $\log_2$ of the hazard ratio - symmetrical for increased/decreased risk; N_{eff}: the effective number of affected patients score.

method	**data type**	**#g**	**adj.p**	**adj.p.model**	**AIC**	**grSize**	**refSize**	**fullSize**	**absHR**	**log2HR**	N_{eff}
ViLoN	int	5	<10^-7	<10^-5	308	30	34	145	51.3	5.7	10401
LRAcluster	int	3	<10^-7	<10^-6	314	42	66	145	24.3	4.6	6734
SNF	int	8	<10^-4	<10^-4	311	17	22	145	46.2	-5.5	5300
clinical	*high_risk*	*2*	*<10^-11*	*<10^-11*	*305.7*	*48*	*97*	*145*	*11.72*	*3.6*	*3549*

With our integrative ViLoN algorithm we can stratify patients into clinically relevant groups significantly better than the alternative integrative approaches already for 2 clusters (Tab. 10), with integrated models for 5 clusters performing even better and best overall (Tab. 11). Our best 2-cluster model (Tab. 10) is already also even better than the 'high_risk' clinical label (Tab. 11).

Compared to the original clinical grouping – high / low risk labels, the ViLoN-established classification is clinically corroborated, *i.e.* high-risk patients are strongly enriched in our groups (Chi-square test *p*-val << 0.001). The 2-cluster grouping looks very close to the clinical labelling but the newly proposed "high-risk" group is extended (Fig. 21, *left-hand* side). The 3-cluster results seem to have interesting internal clusters (Fig. 21, *right-hand* side), with group 1 obviously of highest risk, where the majority of these patients are in fact deceased, group 2 seemingly a moderate risk group, and group 3 comprising of only time-censored patients and with best predicted outcome (Fig. 22).

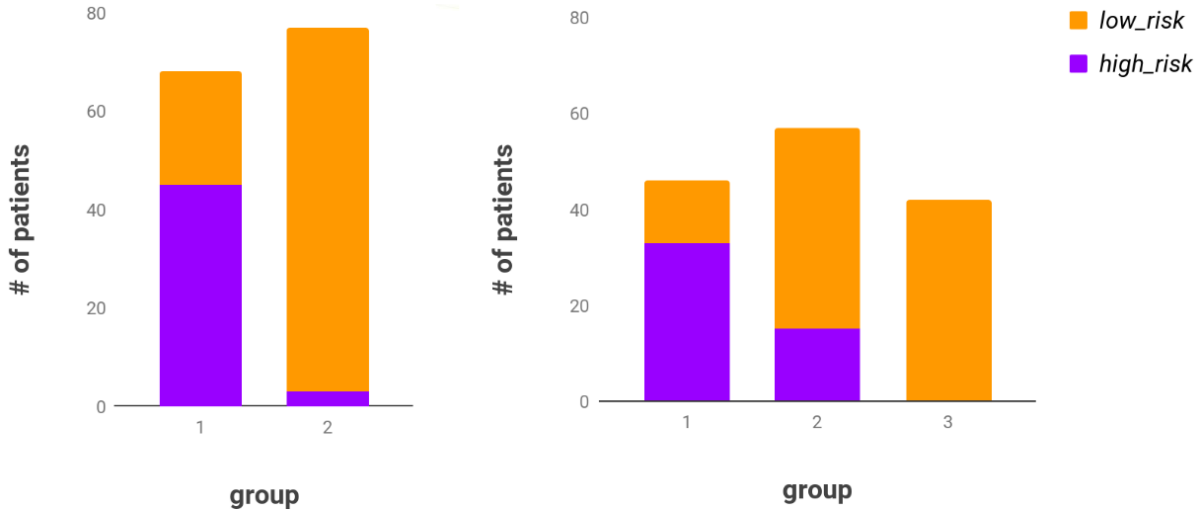

Figure 21: **Bar plots showing the overlap of ViLoN-based groups with the clinical high / low risk labels**. Colors represent different clinical risk labels, with orange color symbolizing 'low risk' and purple color 'high risk' patients. Compared to the original clinical high / low risk labels, the ViLoN-established classification is clinically corroborated. The high-risk patients are strongly enriched in ViLoN groups (Chi-square test p-val << 0.001). The 2-cluster grouping (left-hand side) looks very close to the clinical labelling but the newly proposed 'high-risk' group is extended. The 3-cluster results (right-hand side) yield group 1 of highest risk, where the majority of these patients are in fact deceased (not shown); group 2 of seemingly moderate risk; and group 3 comprising of only time-censored patients and with best predicted outcome.

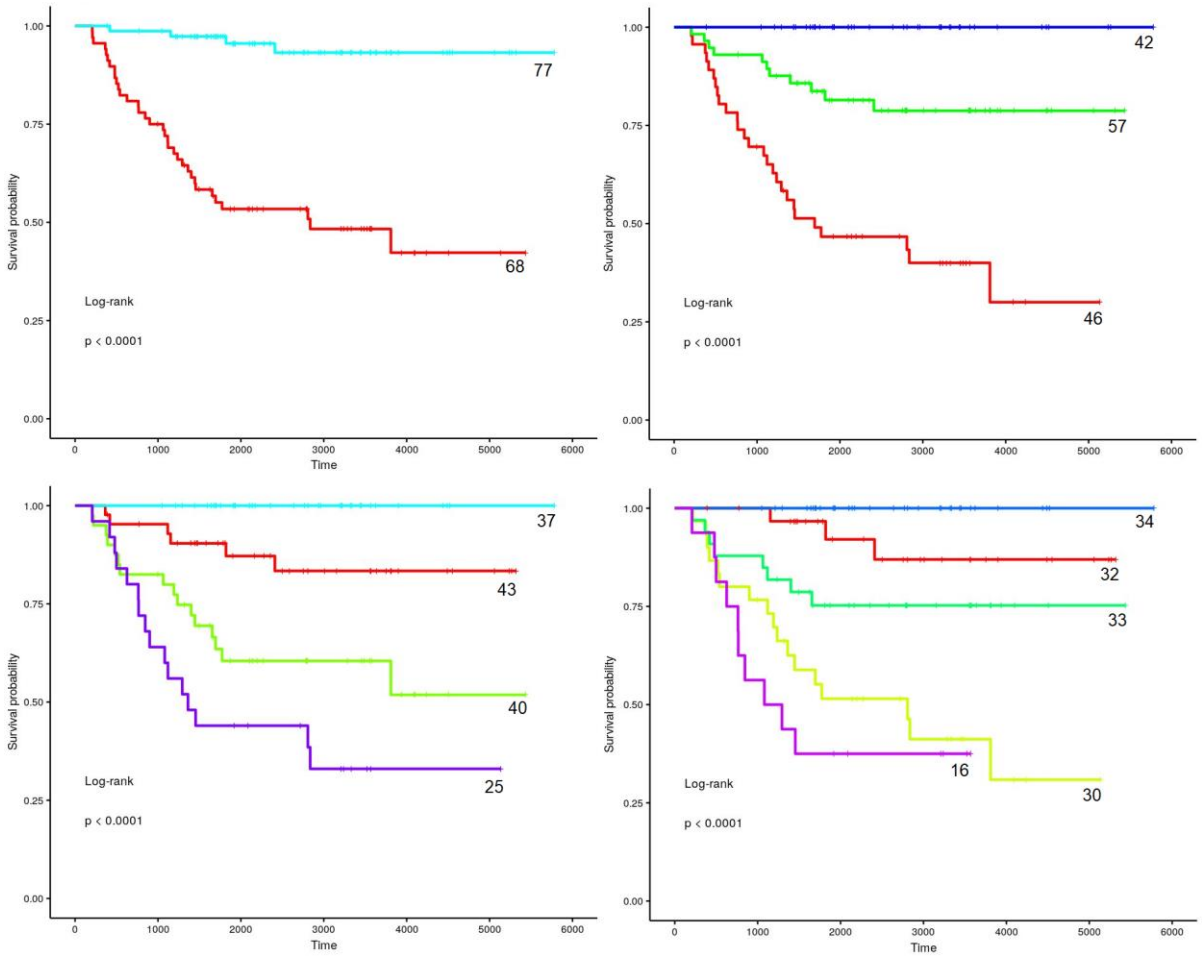

Figure 22: **Kaplan–Meier curves showing the distinct survival profiles of groups found by the ViLoN algorithm in the Neuroblastoma dataset**. Survival profiles of the patient groups found by ViLoN are very well separated, looking remarkably distinct. Notably, these groupings are based solely on the (integrated) molecular data, *i.e.* the network does not have access to the clinical information while grouping the patients. Shown for N=2..5.

5.1.5. Validation on other cancer datasets

In order to further investigate the clinical relevance of our approach we now focus on a well established and largest collection of cancer data. The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), collects data of over 30 types of cancer. It is a large scale collection of molecular data sets profiling patients with various technologies, like, RNA-Seq, methylation, or copy number information[28,98]. The TCGA datasets are publicly available and most of them with a corresponding high-impact publication where each cancer is examined with state-of-the-art methodology. This ensures a well-established baseline for any follow-up research for improving the prevention, diagnosis, and treatment of cancer. To emphasize the potential clinical impact of our novel method we focus on a few cancers where stratification for treatment is known to be difficult. We here discuss one of them and four other cancers are discussed in the Supplement.

Each of the cancer datasets contains matched profiles of two molecular profile types, specifically: RNA-Seq and copy number or methylation data, together with corresponding clinical information. This allows us to build ViLoN networks for exploring integrated analysis. For each resulting integrated network, we stratify patients into *K* subgroups for survival analysis. We then compare ViLoN groupings with models returned by the two integrative methods − SNF and LRAcluster, and seminal models developed for specific TCGA cancers.

We first show an example of integrating gene expression with copy number data, data types similar to the neuroblastoma dataset. In order to show that ViLoN's performance does not depend on the specific molecular profile types available for the neuroblastoma dataset we then focus on integrating gene expression with methylation data. ViLoN further showed good performance on other cancers as well (data not shown).

5.1.5.1. TCGA colon and rectal cancer

At the time of the comprehensive analysis of the human colon and rectal cancer (COADREAD-TCGA) by the TCGA consortium there was no survival information available yet and survival analysis could not be performed[387]. Thus, for reference, we have selected the most recent analysis of the COADREAD dataset[388]. The 'risk score' proposed by Zuo *et al* finds a highly clinically relevant grouping of the patients from this specific dataset (Fig. 23). This is not surprising as the result is adjusted on the whole dataset and is not validated on an independent data. One can thus argue that the yielded product is the highest overall that could be obtained. ViLoN on the other hand clusters patients *de novo*, *i.e.* in an unsupervised manner, and still finds a very relevant grouping. Notably, none of the alternative integrative methods found any significant stratification models.

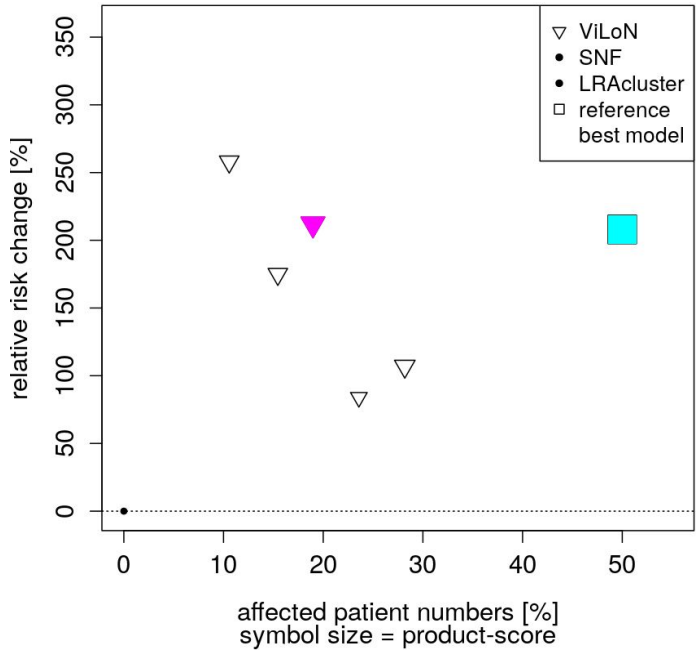

Figure 23: **Comparison of effective number of affected patients scores, in integrative analysis of colon and rectal cancer, with *N*-group stratification.** We compare ViLoN (triangle, point down) with LRAcluster (circle) and SNF (triangle, point up) algorithms and the best TCGA model (square). Shown are all the significant models for *N* between 2 and 9, independent of *N*. The larger the plotted symbol the larger the product, with colors accenting the best achieved scores by each method.

5.1.5.2. TCGA thyroid carcinoma

One of the largest multitrack datasets – thyroid carcinoma (THCA-TCGA), consists of 500 patients for all of whom RNA-Seq and matched methylation samples are provided. It is the most common endocrine malignancy, where identifying high-risk patients for adequate treatment and monitoring remains the major challenge[389]. Specifically, in the case of THCA we compared the resulting products of hazard ratio and group size with the clinical risk assessment stratification proposed in the principal THCA analysis work[390] (Fig. 24). SNF[46] and LRAcluster[42] were also ran on the data, with the latter yielding no significant models. The best performing TCGA model[390] found a group with extremely high hazard ratio, however, for a group of only 27 patients. Interestingly, ViLoN recovered 10 of these patients and found 127 additional ones, still yielding a very high risk of death. Best SNF model was in agreement with ViLoN in case of 33 samples but only for 5 found by the best TCGA model.

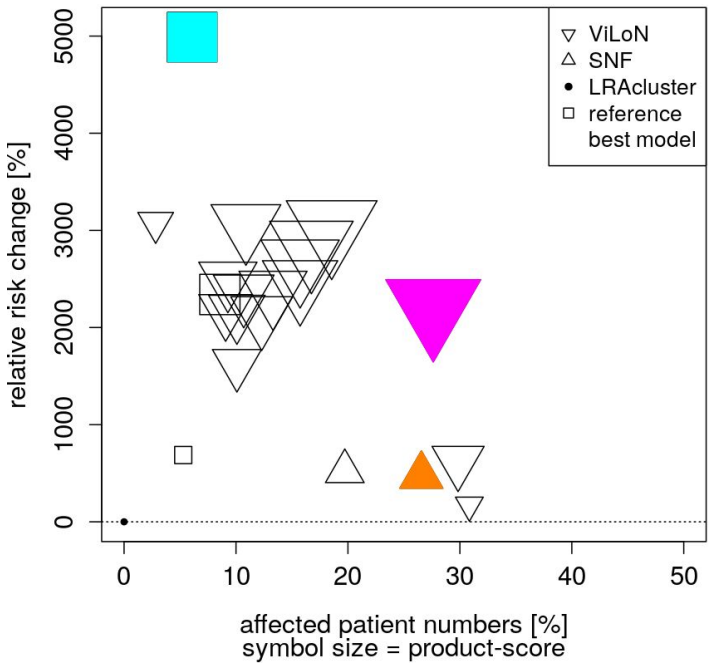

Figure 24: **Comparison of effective number of affected patients scores, in integrative analysis of thyroid carcinoma, with *N*-group stratification.** We compare ViLoN (triangle, point down) with LRAcluster (circle) and SNF (triangle, point up) algorithms and the best TCGA model (square). Shown are all the significant models for *N* between 2 and 9, independent of *N*. The larger the plotted symbol the larger the product, with colors accenting the best achieved scores by each method.

5.1.6. Achieved improvements in patient-stratification and recent ongoing work

I have developed and successfully validated the ViLoN: integrative Variation of Information fused Layers of Networks, algorithm. Our novel method facilitates building a multi-layered integrated network of patients. It consists of multiple single-track patient networks, build with various molecular data profiles. Each patient herself is a network of molecular pathways, build from differential effect data, combined with external biomolecular knowledge. Further, each pathway is itself a network of genes.

I have shown a first application of the ViLoN method to patient stratification and survival analysis of multiple cancers, developing a metric to compare results between tools and publications. Best models were indeed achieved from combination of complementary information: multiple molecular data types and domain knowledge, where ViLoN substantially outperformed alternative integrative methods, single-track analyses, and best models reported in literature.

Stratification into clinical treatment groups based on specific conditions or genes is not always possible, *e.g.*, cells in some cancers are too heterogeneous to be stratified significantly[21,391,392]. In case of the INRG classification system in neuroblastoma, for example, some of the prognostic markers were identified over 30 years ago at a time when no comprehensive approaches for whole genome analysis were available[18]. These are not up-to-date anymore. The presented approach, on the other hand, has been implemented as a scientific workflow that allows regular updates and extensions of results by practitioners in the community.

I am currently also applying the developed networks to advancing statistical multi-layer basic research and also analyse the within-cluster network structure in order to better understand the biology of cancer. The current framework can naturally be extended to include alternative pathway databases, like, Reactome[393] instead of KEGG, and applied to other datasets. We're currently also comparing predictive strength of ViLoN to other newly developed network-based models developed in-house.

5.2. A novel integrative framework for predicting survival time in cancer studies

Even though our ViLoN integrative method could stratify patients into clinically meaningful groups, networks built in this way yielded less robust results when applied to predicting patient survival time (data not shown) – another challenging problem in the research of cancer. In a different approach to data integration[61], we now focus on extracting information that is directly functionally relevant to the new task, using multiple molecular data types from the neuroblastoma dataset[367,368,370,371] and the unusually large METABRIC breast cancer cohort[16].

5.2.1. The importance of vertical and horizontal data integration

As discussed in previous chapters, the heterogeneity of data makes any integrative analysis highly challenging. Data generated with different technologies include different sets of attributes. Where data are highly heterogeneous and weakly related, two interconnected integrative approaches are applied: horizontal and vertical integration (Fig. 25). The horizontal data integration unites information of the same type, but from different data sources and, potentially, in different formats. It facilitates uniting heterogeneous data, like clinical information, from many different sources in one data model. The vertical data integration, on the other hand, means relating different analyses and knowledge across multiple types of data, helping to manage links between the patient's gene expression, clinical information, available chemical knowledge, and existing ontologies. Most existing approaches for data integration focus on one type of data or one disease and cannot facilitate cross-type or -disease integration[394,395].

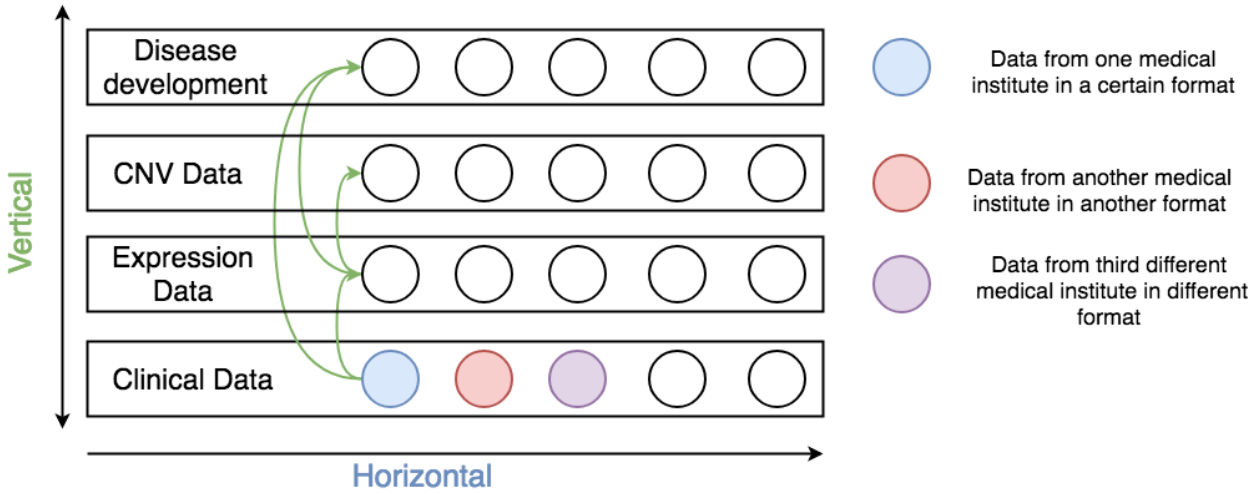

Figure 25: **Horizontal and vertical data integration.** Green arrows show relations between the data types (clinical, expression, CNV and disease development, *i.e.* cancer progression). Horizontal integration is between patients, where the data can originate from, *e.g.*, different institutes, but covers the same type of data. The vertical integration is applied to combine the different data types.

In the following work, horizontal data integration is considered to be a management approach in which the raw data (patients, clinical records, expression profiles, *etc.*) can be "owned" and managed by one network. Usually, each type of raw data can define different semantics for common management purposes. In contrast, vertical integration semantically combines the attributes of each separate type of data that are related to one another. Additional information, in particular for the molecular data, can be found in external domain knowledge sources. With this newly added information the missing parts of the studied data can be filled in. In this way relations between attributes of the different records can be learnt. Currently, there are many established algorithms that address single-track data analyses[371,396–398], and some recent successful approaches to integrative exploration[399]. These, however, usually only focus on one of the integration applications, either horizontal or vertical, underutilizing the entireness of the

available information and the latent relations. We propose a novel framework that employs both these integration views. We show its value on a first example application to machine learning-based survival time prediction.

5.2.2. Combining data from different cancers via dynamic graph databases

Importantly, technical challenges related to data storage and access need to be addressed in order to facilitate streamlined integrative analysis and validation. Data needs to be accessible to an advanced algorithm. Read-only databases, like TCGA[28], that store data for download are essential to moving science forward. In such databases, however, data is related only by patient ids. Moreover, uploading of new patient's information is impossible. Importantly, in such setup all data processing and analysis needs to be performed locally, including development of novel algorithms for data integration. In order to facilitate direct online interaction with the data and to make them accessible to advanced integrative algorithms develop a framework that allows dynamic extension by novel input / output types, and facilitates relating heterogeneous data sources in a meaningful and useful way[61].

Data integration fundamentally involves querying across different data sources. These data sources could be, but are not limited to, separate relational databases or semi-structured data sources distributed across a network. Data integration facilitates dividing the whole data space into two major dimensions, referring to where data or knowledge about metadata reside and to the representation of data and data models. Biomedical experiments take advantage of a vast number of different analytical methods that facilitate mining relevant data from the dispersed information. Some of the most frequent experiments are related to gene expression profiling, clinical data analytics[400], rational drug design[396], which attempt to use all available biological and clinical knowledge to make informed development decisions. Moreover, machine learning-based approaches for finding and highlighting the useful knowledge in the vast space of abundant and heterogeneous data are applied for improving these analytics. Metadata, in particular, are gaining importance, either being captured explicitly or inferred with the help of machine learning models, like, inferring the data structure, data distribution, and common value patterns.

In this study we combine data from neuroblastoma and breast cancer. Via our data integration approach whole datasets are joined, but the semantic integrity of the data is kept and enriched. Through combining data from multiple cancers in this way we create a network of data where entities, like proteins, clinical features and expression features, are linked with each other[401]. Data can be often represented as networks, where nodes indicate biologically relevant entities (typically genes or proteins) and edges represent relationships between these entities (*e.g.*, regulation, interaction). In our overall network that we generate, ultimately, nodes represent patients and edges represent similarities between the patients' profiles, consisting of clinical data, expression profiles and copy number information. Such network can be used to group similar patients and to associate these groups with distinct features[402]. The main challenges here are: (1)

building an appropriate linked data network, discovering a semi-structure of the data model[403] and mapping assertions by the applied model for data integration[404]; and (2) data cleaning, combined into a formal workflow for data integration.

As explained, horizontal data integration means combining data within the same data type (Fig. 25). In the datasets analysed here, these data sources are, specifically: clinical information, expression profiles and copy number data. Each type of data is measured by a different technology and potentially available in various data formats. As an example, we treat clinical data from two cancers as one data source, or one entity, even if they are in different formats. We treat such entities as semantically similar. Vertical data integration, on the other hand, is applied to creating relations between all horizontally integrated objects (Fig. 25). This vertical data integration provides a connection between all the different types of entities. This connection covers relations between patients through clinical information, expression and copy number profiles. Based on these relations we can easily detect all patients closely related to each other by, for instance, protein mutations, diagnosis and/or therapy.

Different databases are required for horizontal and vertical data integration because each of these approaches addresses different aspects of the integration problem. Horizontal data integration deals with unstructured and heterogeneous data. Thus, we use a document-based database (such as MongoDB), which can handle different data types and formats. For vertical data integration a graph-based database is applied, as it is suitable for representing relations. In this study, all relations are established between existing records for each entity, and represented by a semi-structure.

Data integration model with a NoSQL database can potentially unite medical studies data, alternatively to the most frequently used statistical/machine learning methods. Most of the NoSQL database systems share common characteristics, supporting scalability, availability, flexibility and ensuring fast access times for storage, data retrieval and analysis[405,406]. Very often when applying cluster analysis methods for grouping or joining data outliers, small classes, and data that dynamically change relatedness, will cause issues. To overcome these problems a NoSQL database integration model can be applied. Further, we extend the potential of the model by using multiple datasets, regardless of the level of heterogeneity, formats, types of data, *etc.* – all characteristics very relevant in cancer studies[407].

In summary, using the above approach we can relate molecular patient data of different types, and corresponding clinical information, and also combine information between patients and cancers.

Our integrative framework (Fig. 26) facilitates direct analyses of the data. We first focus on a specific clinically relevant application: modeling and prediction of the survival time of cancer patients. This consists of applying both conventional classification methods and machine learning algorithms.

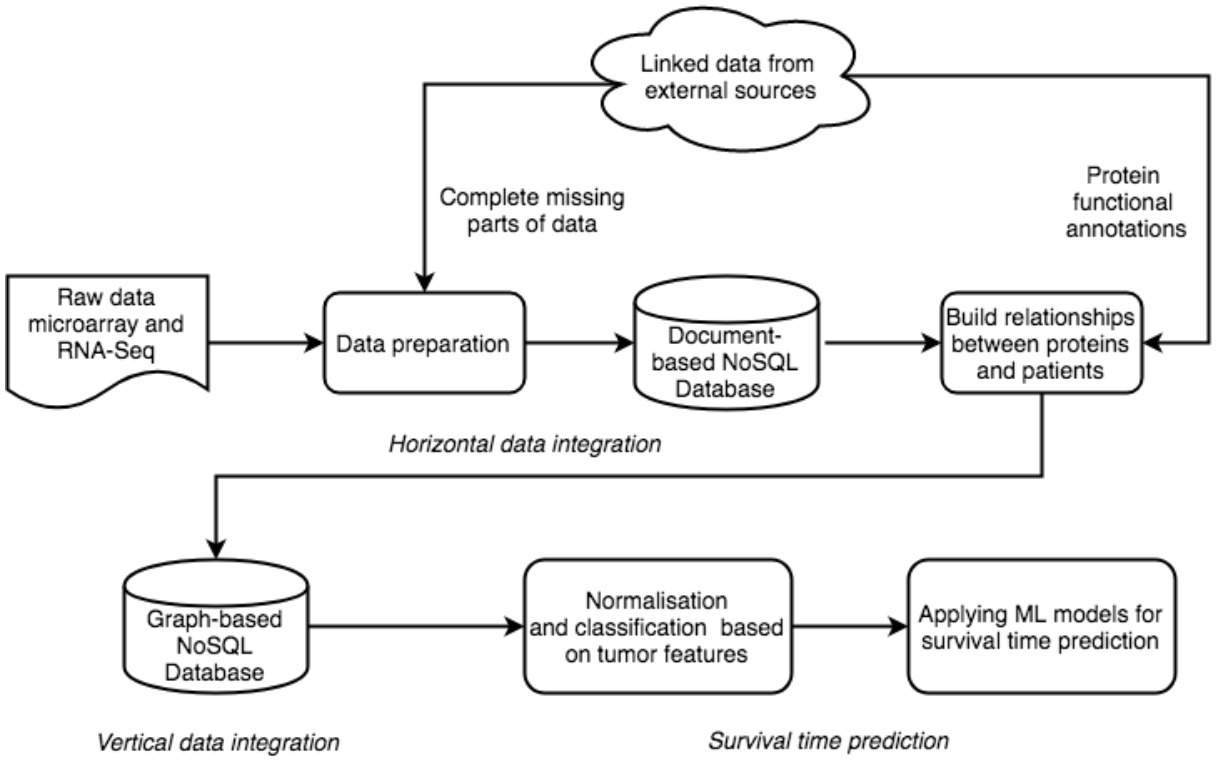

Figure 26: **Workflow of data integration of the independent datasets, performed within our framework.** In data preparation phase we transform and store the raw data of different formats in a document database, performing horizontal data integration per data type. We generate relations between the data based on the available raw patient datasets, including clinical information and molecular data, and we store these in a graph-based database, creating an internal network. We then look up mutated proteins within the networks and search for related information in the external knowledge sources. This way we build the new general relations network which is considered, finally performing the vertical data integration. We store these enriched relations in the graph-based database, together with the internal relationships.

5.2.3. Integrating knowledge about molecular interactions

Through semantic data integration with external domain knowledge sources, like Gene Ontology compendium (GO)[213], UniProt[408], Ensembl[409], we are able to find additional relationships between genes or proteins represented in our data, and extend these to also include other closely related genes/proteins. The strength of relations between these genes is established via a scoring mechanism[213]. Usually, the number of relationships found this way is unfeasibly large (over a billion), substantially increasing the dimensionality of the data. To account for that, we focus only on "trusted relationships". There is no optimal threshold and here we first define "trusted relationships" as relations occurring more than 10 times among different patients. This is necessary for differentiating the significant links between the genes and for reducing the noise of the relationships between the patients through the added gene information. The noise is introduced by the external knowledge sources, where, potentially, all genes can be related. Then we add the relationships found in the external data sources. In general, relations originating from

external knowledge sources are ranked lower, compared to the ones derived from the real datasets.

5.2.4. Linear combination of features for patient clustering

For survival time prediction in breast cancer the Nottingham prognostic index (NPI)[410] is usually applied. It helps to determine prognosis following the surgery. Its value is calculated using three pathological criteria: the size of the lesion, the number of involved lymph nodes, and the grade of the tumor. The NPI can be used to stratify patients into groups and is used to predict five-year survival (in accordance with the more commonly used time scales for survival in other types of cancers)[410]. We do not utilize NPI in our framework because it only applies to one specific disease – breast cancer. In our case a universal predictor is essential, in order to account for other cancers, *e.g.*, neuroblastoma. Thus, we develop a novel and universal predictive parameter – Tumor Integrated Clinical Feature (TICF). To predict patient survival time (in both cancer studies combined) we select specific informative clinical features. We tested different features, their combinations and order, and established the optimal setup (not shown). Specifically, the TICF feature is built by numerically concatenating tumor stage, tumor size and age at diagnosis (Fig. 27) in this exact order. The order of concatenation of the clinical data also shows the importance of clinical information for tumor development and relevance to the patient survival rate. A patient with a tumor in stage four, naturally, will have a shorter survival time compared to patients with a tumor in stage two. The next feature – tumor size, is added second because with an increase of the tumor size the survival rate of a patient is reduced. It is also less important to the survival time than the stage of the tumor. Age at the time of diagnosis, is concatenated third, and indicates that older patients have a lower survival rate. If the order of concatenation of these TICF-composing features would differ patients with distant survival-related features would be incorrectly grouped. In this manner, we provide a normalized distance between patients, essential in our subsequent machine learning approaches to survival time prediction.

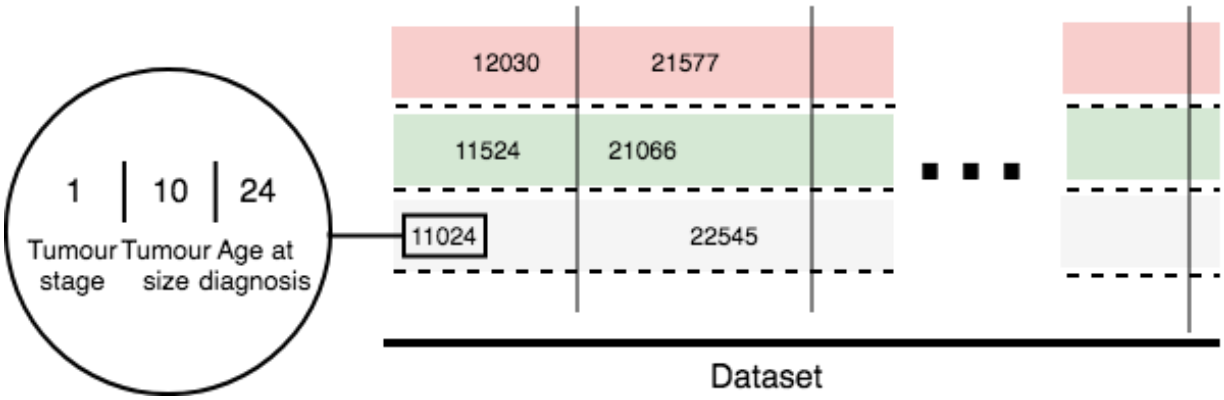

Figure 27: **The universal integrated TICF clinical feature.** TICF consists of three concatenated initial clinical features: tumor stage, tumor size and age at diagnosis. The columns virtually group the patients by TICF, with regard to the first number – the tumor stage. The rows (split by dotted lines) sort patients according to the values of the TICF, referring to the tumor size and age at diagnosis – always from left to right, following the growth of numerical axis.

This novel integrated and universal, *i.e.* applicable to both cancers, feature built by combining clinical features that are most related to survivability can now be used to find patients that are closely related to each other. TICF provides a connection to the linked data network that integrates the molecular information. Specifically, TICF is used in conventional k-neighbours classification, to find the first group of patients that are related most closely to the patient of interest (Fig. 28).

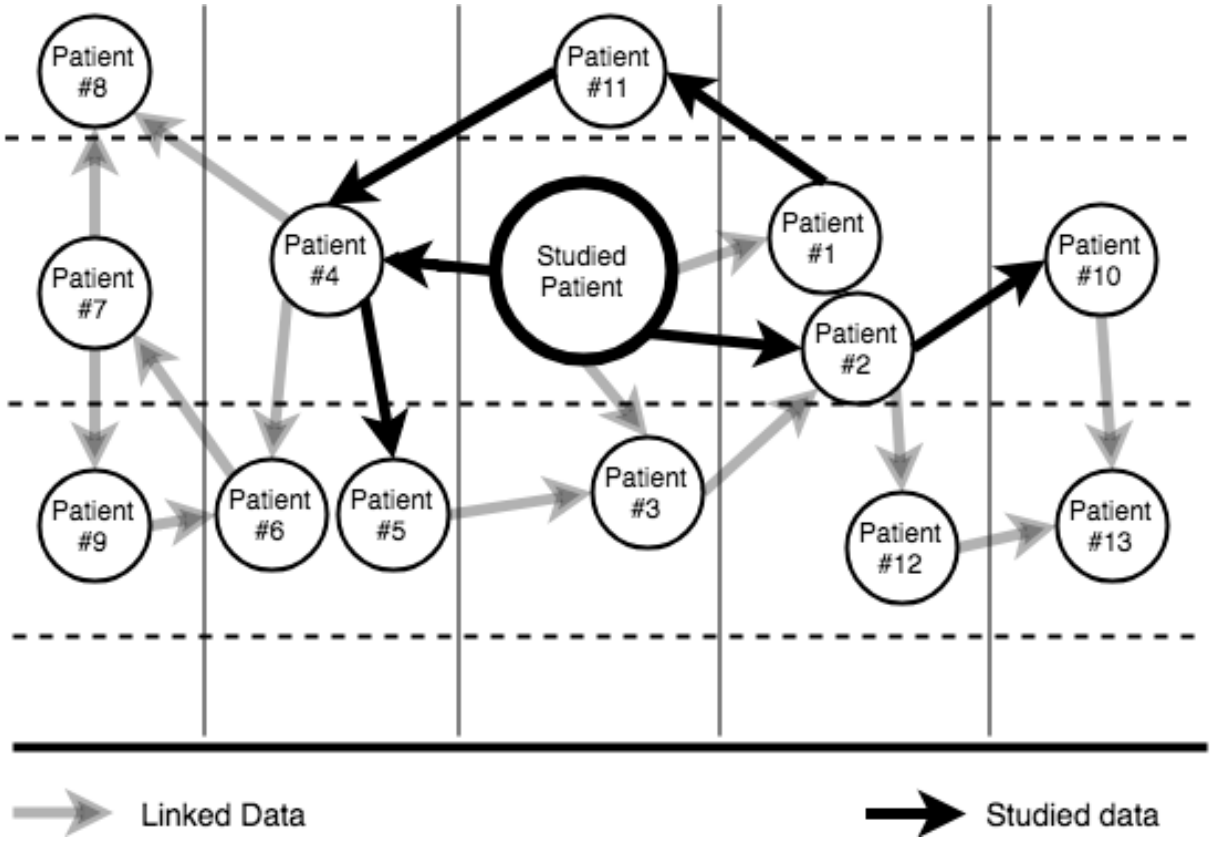

Figure 28: **Example of patients related semantically *via* internal and linked network.** The studied patient of interest is shown as a larger circle in the middle. The vertical lines split the schema into 5 classes, as determined by the k-neighbor classification of the TICF feature. Black arrows show patients related to the studied patient based on the molecular information. Grey arrows distinguish patients related to the studied patient based on external linked data (from EDKS).

After that, we extend this group by finding other patients who may not have the new integrative feature (*e.g.*, due to lack of clinical information) but are still related by different types of molecular data, like gene expression or CNV, od externally introduced information. Machine learning models, based on support vector and decision tree regression, are then used for survival time prediction and cross validation.

5.2.5. Machine learning models for survival time prediction

Next we apply machine learning models to predict and validate survival time of the patients. Artificial intelligence, and in particular machine learning models, has been regularly used in

cancer research, with practical implementations[411]. Artificial neural networks and decision trees, for example, have been used in cancer detection and diagnosis for nearly 30 years[412]. Various models, applying Support Vector Machine (SVM) to cancer prognosis, have been successfully used for approximately two decades[413].

Machine learning models used in our study are based on Support Vector Regression (SVR) with different kernels: Radial Basis Function (RBF), Linear and Poly, and Decision Tree Regression model (DTR). Similar models were shown to perform well for survival prediction in cancer studies[414,415]. Moreover, using these models facilitates a seamless cross-validation.

Specifically, for a patient of interest, for whom we want to predict survival time, we use the extended group of patients, found to be related to her with the TICF feature, and molecular and externally-linked information. These patients have their survival time information available and allow us to build a predictive model with the aforementioned machine learning algorithms, which is then used to infer the survival time of the patient of interest. We compare multiple machine learning methods in a cross-validation approach (Tab. 12), with SVR-Linear and DTR models clearly outperforming other methods.

Table 12: **Aggregated results of cross-validation**: with TICF; with integrative network. The validation is based on four parameters for error evaluation: Train R2 (coefficient of determination) and trained Explained Variance are related to the accuracy of the used model; while trained Negative Mean Absolute Error and Negative Median Absolute Error are related to the noise (error) level.

ML Model	Train R2		Explained Variance		Negative Mean Absolute Error		Negative Median Absolute Error	
	Mean	StD	Mean	StD	Mean	StD	Mean	StD
SVR-RBF	0.318	0.038	0.341	0.032	−45.742	5.224	−34.455	5.594
SVR-LINEAR	0.983	0.007	0.986	0.006	−7.288	1.935	−6.109	1.509
DTR	0.996	0.000	0.996	0.427	−5.624	0.427	−4.636	0.467
SVR-POLY	0.884	0.007	0.887	0.009	−20.354	3.290	−15.581	5.382

To confirm that our integrative framework is indeed necessary to obtain best performance, we examine alternative approaches. First we show cross-validation results using our relational network but without the TICF integrated clinical feature (Tab. 13). Instead we use the clinical features as separate regressors in the experiment. Next we use the TICF feature, but this time we do not extend the patient similarity search with the relational network (Tab. 14). Finally, we look at predictions when only separate clinical features are used and no relational network (Tab. 15). As is evidenced, models building on our novel fully integrative framework outperform the alternatives.

Table 13: **Aggregated results of cross-validation**, **without** the TICF; with integrative network.

ML Model	Train R2		Explained Variance		Negative Mean Absolute Error		Negative Median Absolute Error	
	Mean	StD	Mean	StD	Mean	StD	Mean	StD
SVR-RBF	0.206	0.044	0.230	0.023	−54.923-	5.267	−49.310	8.098
SVR-LINEAR	0.972	0.009	0.973	0.008	−5.494	1.009	−5.312	1.515
DTR	0.991	0.004	0.991	0.004	−5.258	1.125	−4.443	1.235
SVR-POLY	0.883	0.020	0.886	0.021	−18.900	1.112	−14.272	3.289

Table 14: **Aggregated results of cross-validation**, with the TICF; **without** integrative network.

ML Model	Train R2		Explained Variance		Negative Mean Absolute Error		Negative Median Absolute Error	
	Mean	StD	Mean	StD	Mean	StD	Mean	StD
SVR-RBF	0.119	0.171	0.283	0.041	−35.024	3.852	−31.768	8.111
SVR-LINEAR	0.876	0.024	0.896	0.015	−13.733	2.758	−12.726	2.579
DTR	0.984	0.011	0.988	0.006	−4.727	1.760	−4.449	1.709
SVR-POLY	0.006	1.845	0.102	1.708	−21.943	13.774	−14.637	4.944

Table 15: **Aggregated results of cross-validation**, **without** the TICF; **without** integrative network.

ML Model	Train R2		Explained Variance		Negative Mean Absolute Error		Negative Median Absolute Error	
	Mean	StD	Mean	StD	Mean	StD	Mean	StD
SVR-RBF	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
SVR-LINEAR	0.866	0.025	0.877	0.029	−14.485	2.359	−12.594	2.025
DTR	0.985	0.013	0.987	0.011	−4.526	2.133	−4.088	1.684
SVR-POLY	0.764	0.073	0.805	0.033	−17.165	1.376	−14.516	1.963

5.2.6. Building accurate survival time prediction models with data integration

I have shown a novel unified and universal approach for integration of data generated in independent cancer studies. We demonstrated its application to breast cancer and neuroblastoma datasets. Our model is built to facilitate application and extension to multiple different diseases with different types of multi-omics data. Subsequently, we highlight the clinical relevance of our data integration method by applying it to survival time prediction, using machine learning models.

Specifically, we used different database models, with the objective to horizontally and vertically integrate and utilize information, also latent, available in whole and dynamically growing datasets for multiple diseases. Additionally to the extensibility of our data integration model, it also facilitates a seamless integration with external knowledge sources. Specifically, we applied new database technologies: document type database for horizontal integration and graph database for vertical integration – MongoDB and Neo4j, respectively. Such software technology facilitates finding relations between the records in the integrated datasets. The main merit of our approach is

that we are able, also dynamically, to add more data and relations. We explore these opportunities by adding new semantically defined relations from the external knowledge sources. Our software platform can be easily extended and supported, and applied in similar research and practical projects.

Moreover, we developed a new classification feature for survival time prediction. The new TICF feature, is an integrated parameter that facilitates finding patients closely related to the patient we want to study, and make predictions for. Inclusion of related patients with different clinical and expression parameters, as we show, is essential for improving the accuracy of survival prediction models.

For survival time prediction we apply supervised regression models. Models used in this study utilize the TICF feature to improve the accuracy of patient survival time prediction. Moreover, application of these specific machine learning algorithms ensures a reliable validation of our semantic data integration approach. Cross-validation of these models showed stable results with regard to achieved performance – both in the context of success and error rates, in survival time prediction.

In order to demonstrate the extensibility of our system we have now extended our framework to facilitate exploration of other cancers – BRCA, KIRC, and LUAD, and potentially all TCGA cancers, adding automated data retrieval of the TCGA / GDC datasets. We also considerably improved the pipeline for model evaluation with improved cross-validation, also extending the number of machine learning algorithms readily available within our framework (data not shown).

6. Chapter 6: Conclusion: Towards precision medicine through data integration

In this thesis I advanced the state-of-the-art in exploiting complex data with ill-understood biases and noise characteristics in the biomedical domain, especially matched molecular profiles of different types and clinical data from cancer patients. While much hope has been placed in such integrative analyses, such work is surprisingly rare (Fig. 1 *right-hand* side, *Chapter 1*). Moreover, and perhaps counterintuitively, such analyses make up a decreasing percentage of published research (Fig. 1 right-*hand* side, *Chapter 1*).

This is likely due to multiple progress-hindering factors related to the variety of technologies used to generate the different types of data and various stages of sample processing. Each data type has different characteristics, with different type-specific random variation, or noise, present, and bias introduced by multiple different factors. Even the noise in data types that should seemingly be very similar, because in essence they are both used to assess gene expression – microarrays and RNA-Seq, will possess very distinct characteristics that need to be accounted for when analysing the data jointly. Microarray data are known to be highly noisy, including noise of

different level and impact on the analysis[169–173], introduced at various stages of the microarray experiment, such as hybridisation and image analysis[174]. Together with systematic biases introduced directly by handling the samples by a human operator, results generated from microarray analysis are not straightforward to reproduce[166,168]. When working with microarray data all these issues need to be accounted for and the data processed and analysed accordingly[167]. Notably, copy number data, based on hybridization of genomic DNA to microarrays, will have similar noise issues[175]. In the case of RNA-Seq, it is known that in the random sampling process that is inherent to the assay, dependent also on the sequencing depth, measurement noise is introduced. In addition, even though hybridization-based noise associated with microarrays is not present[176], background noise from genomic DNA contamination might be[94]. In RNA-seq data also a transcript length bias needs to be addressed, where a strong association exists between the length of the transcript and the ability of a differential effect analysis method to call genes differentially expressed between samples[177]. Again, all these can and must be accounted for when processing and analysing the data. Assuming that the noise and biases inherent to specific technology are controlled for, another potential bias might be introduced at the stage of joint analysis, as different technological platforms assay different numbers of genes. Here a solution is to focus the analysis on genes measured by both platforms[416]. In fact, there is an understanding that combining data generated with different technologies may offset and compensate for the different technical artefacts or bias specific to a measurement platform, eventually leading to more robust biologically relevant readouts[29,417].

In order to advance the state-of-the-art in exploiting complex data, I have developed and validated a novel integrative framework – *Variation of information fused Layers of Networks* (ViLoN), with a first application to cancer patient stratification into treatment groups. Cancer patients are represented in a multi-layer network view, combining matched molecular profiles and external knowledge about biochemical pathways (KEGG) and molecular processes (GO terms) (Fig. 18A, *Chapter 5*). This approach reduces the dimensionality by summarizing molecular data, creating a pathway level view. In this way we characterize the disease of a specific patient as reflected in a particular molecular profile type and already incorporating functional knowledge. After data preprocessing, this is the crucial first step in my integrative framework, helping to overcome the *curse of dimensionality*, that otherwise limits the interpretation of biological signal, with hundreds to thousands of variables, *i.e.* genes, measured but for relatively small number of samples[164,165]. This way we summarize the patient-specific molecular effects at pathway level[11], in relation to the rest of the patients in the cohort. This is, moreover, a form of data integration already, joining prior domain biochemical knowledge with patient-specific gene expression, combining measurements across genes. We then further reduce the dimensionality of the network by clustering the pathways (per patient). This set of pathway clusters can be used to characterize the disease of a specific patient as reflected in a particular molecular profile type and already incorporating functional knowledge. Then, by a information-theoretic measure, a distance between the clusterings is found, indicating patient similarities. We construct a robust patient similarity graph (Fig. 18B, *Chapter 5*) for each molecular profile type

(like, gene expression, copy number, *etc.*), to finally combine information across different molecular profile types, yielding a single integrated patient similarity graph (Fig. 18C, *Chapter 5*). This integrated network is then exploited by a clustering algorithm to classify individuals into groups of patients that should potentially be treated clinically in a similar fashion.

Critically, in order to objectively demonstrate that a new method is indeed of avail, it needs to be assessed through effective benchmarks, *e.g.*, *via* gold-standards, in comparison with established alternatives[49–51]. While the development of improved approaches to data integration constitutes an active field, there is an open need for effective benchmarks of algorithm performance in general. Selecting a method amongst available alternatives has thus remained a non-trivial challenge for practitioners in biomedical research. Moreover, because different studies focus on different objectives it is difficult to compare methods that intrinsically focus on answering different questions and yield results that are not directly comparable with other studies. Multiple factors, specific to the problem being answered, need to be considered. For instance, for a meaningful evaluation of a method for patient stratification, we need to consider not only the significance of the novel grouping but also the number of patients and the size of the risk change. Furthermore, it is noteworthy that without a negative reference set providing known or expected negatives, it is impossible to estimate the specificity of a method, which is based on the number of false positives that a method calls wrongly. For example, when evaluating a novel method for gene enrichment analysis, while scores that assess how highly known positives are ranked do provide objective measures of sensitivity[59,418,419], they cannot discriminate methods that identify potentially valuable additional uncharacterized pathways from methods that wrongly include known or expected negatives. Consequently, exploiting known negative examples has considerably improved method performance in various areas of computational biology, such as the prediction of mRNA targets of regulatory microRNAs[54] or the prediction of protein function[55], *e.g.*, for the identification of DNA-binding proteins from their sequences[56]. Moreover, because of the difficulty of determining negative samples for benchmarks, researchers seek to exploit simulated data where some ground truth is known[57,58]. For simulations to be of relevance, however, they need to capture the biological complexity and internal structure of datasets obtained under realistic conditions[12,14,59,60] – yet another non-trivial challenge.

To assess the performance of our newly developed algorithm we have established a novel rigorous and balanced metric – *effective number of affected patients* (N_{eff}), dedicated to validating the performance of a method for patient stratification. Our metric accounts for the clinical relevance of the grouping, recognizing not only the size of the effect of reduced or increased risk of death, *i.e.* hazard ratio, but also considering the number of affected patients. It facilitates direct assessment of clinical relevance of patient stratification into distinct groups for treatment, comparing multiple state-of-the-art tools. Notably, our metric also allows comparisons of results presented in independent scientific reports. This was previously not possible because the usual metrics presented in scientific publications are not directly comparable with each other, yielded for various group sizes, not directly accounting for clinical impact.

Intriguingly, we achieved the best performing models indeed with combination of complementary information, integrating multiple molecular data types per patient with external domain knowledge about structure in the data, such as pathways. The difference in performance between single molecular data type based models is substantial and our novel framework outperforms both the results shown in relevant literature (Tab. 9, *Chapter 5*) and alternative integrative state-of-the-art algorithms (Tab. 10, *Chapter 5*). Remarkably, our method builds solely on molecular information and external domain knowledge, not considering clinical information for the predictions.

Even though our ViLoN integrative method performed remarkably well for patient stratification, it yielded less robust results for predicting patient survival time – another challenging problem in the research of cancer. Therefore, in a different approach to data integration, we focus on extracting information that is directly functionally relevant to the new task. In order to ensure seamless and dynamic access to the data, in a way that would also allow for inferring vertical and horizontal relations between different data types and patients, and facilitate complex analyses, we developed a complete semantic data integration system[61]. In our framework we relate patients by the similarity of their heterogeneous data, first finding a group of patients related to the patient of interest by the means of the KNN machine learning algorithm applied to the novel integrated clinical feature we developed (Fig. 27, *Chapter 5*). We then extend such a group to include patients that further relate to the patients already included in the group, by the similarity of their molecular data (Fig. 28, *Chapter 5*). Survival time for the specific patient can then be seamlessly predicted based on the relevant clinical information of the patients that are most similar to the studied patient of interest. In this way we explore molecular effect, like gene expression, not as a direct predictor, which is known to be extremely difficult because of the high dimensionality of the data[420], but for finding patients similar to the one whose survival time we want to predict. Satisfyingly, the best models are now achieved with combination of complementary information, contained within the integrative clinical feature and the network structure built using the molecular data and externally linked domain knowledge.

6.1. Summary and outlook

In conclusion, in this thesis I have shown that exploration of the notoriously abundant biomolecular data presents scientists with a great challenge, at the same time giving us a great opportunity for facilitating an advancement in our understanding of biology in general and human disease in particular. I have shown that these Big Data are not yet fully utilized because of their high dimensionality and heterogeneity. I have further presented my contributions to the advancement of exploiting these complex data, showing that by reducing the dimensionality of molecular data we can extract higher-level actionable patterns, and attenuate the inherent noise. Specifically, I have developed novel cutting edge integrative frameworks, not only being able to join high dimensional and heterogeneous information generated by multiple technologies, but also share the information across patients. In rigorous validation procedures I have shown the

potential of our methods to improve and target treatment of cancer patients. I hope that my work will contribute to further advancement of precision medicine.

With my integrative algorithms I am currently exploring other cancer datasets, with notably over thirty being hosted by the TCGA database. Even though, the method for patient stratification I have proposed already achieved substantial improvement over state-of-the-art, it possibly still can be further optimized. KEGG pathway database could be exchanged for a much larger Reactome[393] pathway database. This would be a more demanding task computationally, because the patient graphs generated with such large data, would be significantly larger than graphs generated using the much smaller KEGG pathway database. Moreover, my tool could be extended to apply alternative clustering algorithms which could potentially yield different patient stratifications. Using the networks generated in this work I am also currently working on a multi-layer network method for finding most dysregulated pathways between cancer patient groups. With regard to the integrative framework for survival analysis, in addition to analysing further TCGA cancer datasets, I am now exploring novel machine learning methods for the task.

7. Bibliography

1. Ribitsch, I. *et al.* Fetal articular cartilage regeneration versus adult fibrocartilaginous repair: secretome proteomics unravels molecular mechanisms in an ovine model. *Dis. Model. Mech.* **11,** (2018).

2. Dame, K. *et al.* Thyroid Progenitors Are Robustly Derived from Embryonic Stem Cells through Transient, Developmental Stage-Specific Overexpression of Nkx2-1. *Stem Cell Reports* **8,** 216–225 (2017).

3. Benton, D. Bioinformatics--principles and potential of a new multidisciplinary tool. *Trends Biotechnol.* **14,** 261–272 (1996).

4. Mushegian, A. Grand challenges in bioinformatics and computational biology. *Front. Genet.* **2,** 60 (2011).

5. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).

6. Martínez, E. *et al.* Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects. *Oncogene* **34,** 2732–2740 (2015).

7. Kobayashi, T. *et al.* Microarray analysis of gene expression at the tumor front of colon cancer. *Anticancer Res.* **35,** 6577–6581 (2015).

8. Numata, K. *et al.* Clinical significance of IGF1R gene expression in patients with Stage II/III gastric cancer who receive curative surgery and adjuvant chemotherapy with S-1. *J. Cancer Res. Clin. Oncol.* **142,** 415–422 (2016).

9. Huang, T., Wu, W., Jin, H. & Cai, Y.-D. Gene Sets of Gene Ontology are More Stable Diagnostic Biomarkers than Genes in Oral Squamous Cell Carcinoma. *Curr. Bioinform.* **8,** 577–582 (2013).

10. Kim, H., Watkinson, J. & Anastassiou, D. Biomarker discovery using statistically significant gene sets. *J. Comput. Biol.* **18,** 1329–1338 (2011).

11. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102,** 15545–15550 (2005).

12. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinformatics* **13,** 281–291 (2012).

13. Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J. W. L. & Haibe-Kains, B. Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.* **4,** 4092 (2014).

14. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8,** e1002375 (2012).

15. Bossi, P. *et al.* Functional Genomics Uncover the Biology behind the Responsiveness of Head and Neck Squamous Cell Cancer Patients to Cetuximab. *Clin. Cancer Res.* **22,** 3961–3970 (2016).

16. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel

subgroups. *Nature* **486,** 346–352 (2012).

17. Papaemmanuil, E. *et al.* Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374,** 2209–2221 (2016).

18. Bagatell, R. & Cohn, S. L. Genetic discoveries and treatment advances in neuroblastoma. *Curr. Opin. Pediatr.* **28,** 19–25 (2016).

19. Oberthuer, A. *et al.* Subclassification and individual survival time prediction from gene expression data of neuroblastoma patients by using CASPAR. *Clin. Cancer Res.* **14,** 6590–6601 (2008).

20. Cohn, S. L. *et al.* The International Neuroblastoma Risk Group (INRG) classification system: an INRG Task Force report. *J. Clin. Oncol.* **27,** 289–297 (2009).

21. Fardin, P. *et al.* A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients. *Mol. Cancer* **9,** 185 (2010).

22. Milioli, H. H. *et al.* Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. *BioData Min.* **9,** 2 (2016).

23. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* **99,** 6567–6572 (2002).

24. Mueller, S. & Matthay, K. K. Neuroblastoma: biology and staging. *Curr. Oncol. Rep.* **11,** 431–438 (2009).

25. Fagerholm, R. *et al.* The SNP rs6500843 in 16p13.3 is associated with survival specifically among chemotherapy-treated breast cancer patients. *Oncotarget* **6,** 7390–7407 (2015).

26. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).

27. McDermott, U. Next-generation sequencing and empowering personalised cancer medicine. *Drug Discov. Today* **20,** 1470–1475 (2015).

28. TCGA Research Network. The Cancer Genome Atlas. at <https://www.cancer.gov/tcga>

29. Nguyen, T., Diaz, D., Tagett, R. & Draghici, S. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Sci. Rep.* **6,** 29251 (2016).

30. Nguyen, T., Tagett, R., Donato, M., Mitrea, C. & Draghici, S. A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics* **32,** 409–416 (2016).

31. Nguyen, T. C. Horizontal And Vertical Integration Of Bio-Molecular Data. (2017).

32. Searls, D. B. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* **4,** 45–58 (2005).

33. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems* **96,** 86–103 (2009).

34. Yuan, Y., Savage, R. S. & Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.* **7,** e1002227 (2011).

35. Schlicker, A., Michaut, M., Rahman, R. & Wessels, L. F. A. OncoScape: Exploring the cancer aberration landscape by genomic data fusion. *Sci. Rep.* **6,** 28103 (2016).

36. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912 (2009).

37. Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7,** e35236 (2012).

38. Kuijjer, M. L. *et al.* Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. *Genes Chromosomes Cancer* **51,** 696–706 (2012).

39. Louhimo, R., Lepikhova, T., Monni, O. & Hautaniemi, S. Comparative analysis of algorithms for integration of copy number and expression data. *Nat. Methods* **9,** 351–355 (2012).

40. Lu, T.-P. *et al.* Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PLoS ONE* **6,** e24829 (2011).

41. Ahmed, A. *et al.* The conundrum of diagnosing cutaneous composite lymphoma in the molecular age. *Am. J. Dermatopathol.* **41,** 757–766 (2019).

42. Wu, D., Wang, D., Zhang, M. Q. & Gu, J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* **16,** 1022 (2015).

43. Pham, L., Christadore, L., Schaus, S. & Kolaczyk, E. D. Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proc Natl Acad Sci USA* **108,** 13347–13352 (2011).

44. Verbeke, L. P. C. *et al.* Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS ONE* **10,** e0133503 (2015).

45. Mizrachi, E. *et al.* Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proc Natl Acad Sci USA* **114,** 1195–1200 (2017).

46. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11,** 333–337 (2014).

47. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

48. Hamid, J. S. *et al.* Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics* **2009,** (2009).

49. Aniba, M. R., Poch, O. & Thompson, J. D. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res.* **38,** 7353–7363 (2010).

50. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14,** 1063–1071 (2017).

51. Boutros, P. C., Margolin, A. A., Stuart, J. M., Califano, A. & Stolovitzky, G. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol.* **15,** 462 (2014).

52. Kühberger, A., Fritz, A., Lermer, E. & Scherndl, T. The significance fallacy in inferential statistics.

BMC Res. Notes **8,** 84 (2015).

53. Azuero, A. A note on the magnitude of hazard ratios. *Cancer* **122,** 1298–1299 (2016).

54. Bandyopadhyay, S. & Mitra, R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* **25,** 2625–2631 (2009).

55. Youngs, N., Penfold-Brown, D., Bonneau, R. & Shasha, D. Negative example selection for protein function prediction: the NoGO database. *PLoS Comput. Biol.* **10,** e1003644 (2014).

56. Xu, R. *et al.* enDNA-Prot: identification of DNA-binding proteins by applying ensemble learning. *Biomed Res. Int.* **2014,** 294279 (2014).

57. Wei, Z. & Li, H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* **8,** 265–284 (2007).

58. Tyekucheva, S., Marchionni, L., Karchin, R. & Parmigiani, G. Integrating diverse genomic data using gene sets. *Genome Biol.* **12,** R105 (2011).

59. Clark, N. R. *et al.* Principal angle enrichment analysis (PAEA): dimensionally reduced multivariate gene set enrichment analysis tool. *Proceedings (IEEE Int Conf Bioinformatics Biomed)* **2015,** 256–262 (2015).

60. Benidt, S. & Nettleton, D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics* **31,** 2131–2140 (2015).

61. Mihaylov, I., Kańduła, M., Krachunov, M. & Vassilev, D. A Novel Framework for Horizontal and Vertical Data Integration in Cancer Studies with Application to Survival Time Prediction Models. *Biol Direct* (2019).

62. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).

63. Ioannidis, J. P. A. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294,** 218–228 (2005).

64. Ioannidis, J. P. A. How to make more published research true. *PLoS Med.* **11,** e1001747 (2014).

65. Hothorn, T. & Leisch, F. Case studies in reproducibility. *Brief. Bioinformatics* **12,** 288–300 (2011).

66. Hofner, B., Schmid, M. & Edler, L. Reproducible research in statistics: A review and guidelines for the Biometrical Journal. *Biom. J.* **58,** 416–427 (2016).

67. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46,** W537–W544 (2018).

68. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

69. Lampa, S., Alvarsson, J. & Spjuth, O. Towards agile large-scale predictive modelling in drug discovery with flow-based programming design principles. *J. Cheminform.* **8,** 67 (2016).

70. Sadedin, S. P., Pope, B. & Oshlack, A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28,** 1525–1526 (2012).

71. Tschager, T. & Schmidt, H. A. DAGwoman: Enabling DAGMan-like workflows on non-Condor

platforms. in *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies - SWEET '12* 1–6 (ACM Press, 2012). doi:10.1145/2443416.2443419

72. Wolstencroft, K. *et al.* The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **41,** W557-61 (2013).

73. Spjuth, O. *et al.* Experiences with workflows for automating data-intensive bioinformatics. *Biol. Direct* **10,** 43 (2015).

74. Marx, V. Biology: The big challenges of big data. *Nature* **498,** 255–260 (2013).

75. Bux, M. & Leser, U. Parallelization in Scientific Workflow Management Systems. *ArXiv e-prints* (2013).

76. Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20,** 3045–3054 (2004).

77. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **Chapter 19,** Unit 19.10.1-21 (2010).

78. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15,** 1451–1455 (2005).

79. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).

80. Altintas, I. *et al.* Kepler: an extensible system for design and execution of scientific workflows. in *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.* 423–424 (IEEE, 2004). doi:10.1109/SSDM.2004.1311241

81. Kallio, M. A. *et al.* Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12,** 507 (2011).

82. Feldman, S. I. Make — a program for maintaining computer programs. *Softw: Pract. Exper.* **9,** 255–265 (1979).

83. Schwab, M. & Schroeder, J. Reproducible research documents using GNUmake. in *Stanford Exploration Project* **SEP-89,** 217–226 (1995).

84. Schatz, M. C., Langmead, B. & Salzberg, S. L. Cloud computing and the DNA data race. *Nat. Biotechnol.* **28,** 691–693 (2010).

85. Dean, J. & Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Sixth Symposium on Operating System Design and Implementation: 2004; San Francisco, CA* (2004).

86. White, T. *Hadoop: The Definitive Guide*. (O'Reilly, 2009).

87. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S. & Stoica, I. Spark: cluster computing with working sets. in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing* 10–10 (2010).

88. Lampa, S., Dahlö, M., Olason, P. I., Hagberg, J. & Spjuth, O. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data.

Gigascience **2,** 9 (2013).

89. Rodríguez, D., Bello, X. & Gutiérrez-de-Terán, H. Molecular Modelling of G Protein-Coupled Receptors Through the Web. *Mol. Inform.* **31,** 334–341 (2012).

90. Schönherr, S. *et al.* Cloudgene: a graphical execution platform for MapReduce programs on private and public clouds. *BMC Bioinformatics* **13,** 200 (2012).

91. Siretskiy, A., Sundqvist, T., Voznesenskiy, M. & Spjuth, O. A quantitative assessment of the Hadoop framework for analyzing massively parallel DNA sequencing data. *Gigascience* **4,** 26 (2015).

92. Siretskiy, A. & Spjuth, O. HTSeq-Hadoop: Extending HTSeq for Massively Parallel Sequencing Data Analysis using Hadoop. in *eScience (eScience), 2014 IEEE 10th International Conference on* (2014).

93. Anders, S., Pyl, P. T. & Huber, W. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

94. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32,** 903–914 (2014).

95. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32,** 888–895 (2014).

96. Mueckstein, U., Leparc, G. G., Posekany, A., Hofacker, I. & Kreil, D. P. Hybridization thermodynamics of NimbleGen microarrays. *BMC Bioinformatics* **11,** 35 (2010).

97. Goodstadt, L. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics* **26,** 2778–2779 (2010).

98. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

99. Ikonomou, L. *et al.* The Genetic Program of Primordial Lung Progenitors. *Nat Commun* (2019).

100. Schumacher, A. *et al.* SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics* **30,** 119–120 (2014).

101. Niemenmaa, M. *et al.* Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics* **28,** 876–877 (2012).

102. Merali, Z. Computational science: ...Error. *Nature* **467,** 775–777 (2010).

103. Orrù, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell* **155,** 242–256 (2013).

104. Francalacci, P. *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341,** 565–569 (2013).

105. Cuccuru, G. *et al.* An automated infrastructure to support high-throughput bioinformatics. in *2014 International Conference on High Performance Computing & Simulation (HPCS)* 600–607 (IEEE, 2014). doi:10.1109/HPCSim.2014.6903742

106. Pireddu, L., Leo, S. & Zanetti, G. SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27,** 2159–2160 (2011).

107. Pireddu, L., Leo, S., Soranzo, N. & Zanetti, G. A Hadoop-Galaxy adapter for user-friendly and scalable data-intensive bioinformatics in Galaxy. in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14* 184–191 (ACM Press, 2014). doi:10.1145/2649387.2649429

108. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32,** 25–32 (2011).

109. Weissensteiner, H. *et al.* mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.* **44,** W64-9 (2016).

110. Afgan, E. *et al.* Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics* **11 Suppl 12,** S4 (2010).

111. Forer, L. *et al.* Delivering bioinformatics MapReduce applications in the cloud. in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 373–377 (IEEE, 2014). doi:10.1109/MIPRO.2014.6859593

112. Krachunov, M. Hierarchy and expressions for automated workflows for NGS data processing. in *Proceedings of the 8th International Conference on Information Systems & Grid Technologies (ISGT)* 38–48 (2015).

113. Schaaff, A., Verdes-Montenegro, L., Ruiz, J. E. & Vela, J. S. Scientific workflows in astronomy. *Proceeding of Astronomical Data Analysis Software and Systems* (2012).

114. Lih, A. & Zadok, E. *PGMAKE: A Portable Distributed Make System*. (1994).

115. Taura, K. *et al.* Design and implementation of GXP make — A workflow system based on make. *Future Generation Computer Systems* **29,** 662–672 (2013).

116. Albrecht, M., Donnelly, P., Bui, P. & Thain, D. Makeflow: A portable abstraction for data intensive computing on clusters, clouds, and grids. in *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies - SWEET '12* 1–13 (ACM Press, 2012). doi:10.1145/2443416.2443417

117. Seibel, P. N. *et al.* XML schemas for common bioinformatic data types and their application in workflow systems. *BMC Bioinformatics* **7,** 490 (2006).

118. Wilkinson, M. Interoperability With Moby 1.0 - It's Better Than Sharing Your Toothbrush! *Available from Nature Precedings* (2008).

119. Linke, B., Giegerich, R. & Goesmann, A. Conveyor: a workflow engine for bioinformatic analyses. *Bioinformatics* **27,** 903–911 (2011).

120. Wassink, I. *et al.* Analysing Scientific Workflows: Why Workflows Not Only Connect Web Services. in *Services - I, 2009 World Conference on* 314–321 (2009).

121. Sayegh, R. G. *et al.* Polarization-Sensitive Optical Coherence Tomography and Conventional Retinal Imaging Strategies in Assessing Foveal Integrity in Geographic Atrophy. *Invest. Ophthalmol. Vis. Sci.* **56,** 5246–5255 (2015).

122. Gaignebet, L. *et al.* Sex-specific human cardiomyocyte gene regulation in left ventricular pressure overload. *Mayo Clin. Proc* (2019).

123. Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F. & Munafò, M. R. Low statistical power in biomedical science: a review of three human research domains. *R. Soc. Open Sci.* **4,** 160254 (2017).

124. Regitz-Zagrosek, V. & Kararigas, G. Mechanistic pathways of sex differences in cardiovascular disease. *Physiol. Rev.* **97,** 1–37 (2017).

125. Carroll, J. D. *et al.* Sex-associated differences in left ventricular function in aortic stenosis of the elderly. *Circulation* **86,** 1099–1107 (1992).

126. Villar, A. V. *et al.* Gender differences of echocardiographic and gene expression patterns in human pressure overload left ventricular hypertrophy. *J. Mol. Cell. Cardiol.* **46,** 526–535 (2009).

127. Kararigas, G. *et al.* Sex-dependent regulation of fibrosis and inflammation in human left ventricular remodelling under pressure overload. *Eur. J. Heart Fail.* **16,** 1160–1167 (2014).

128. Weinberg, E. O. *et al.* Gender differences in molecular remodeling in pressure overload hypertrophy. *J Am Coll Cardiol* **34,** 264–273 (1999).

129. Fliegner, D. *et al.* Female sex and estrogen receptor-beta attenuate cardiac remodeling and apoptosis in pressure overload. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **298,** R1597-606 (2010).

130. Queirós, A. M. *et al.* Sex- and estrogen-dependent regulation of a miRNA network in the healthy and hypertrophied heart. *Int. J. Cardiol.* **169,** 331–338 (2013).

131. Kararigas, G. *et al.* Comparative proteomic analysis reveals sex and estrogen receptor β effects in the pressure overloaded heart. *J. Proteome Res.* **13,** 5829–5836 (2014).

132. Sanchez-Ruderisch, H. *et al.* Sex-specific regulation of cardiac microRNAs targeting mitochondrial proteins in pressure overload. *Biol. Sex Differ.* **10,** 8 (2019).

133. Gladka, M. M. *et al.* Single-Cell Sequencing of the Healthy and Diseased Heart Reveals Cytoskeleton-Associated Protein 4 as a New Modulator of Fibroblasts Activation. *Circulation* **138,** 166–180 (2018).

134. Douglas, P. S. *et al.* Gender differences in left ventricle geometry and function in patients undergoing balloon dilatation of the aortic valve for isolated aortic stenosis. *Nhlbi balloon valvuloplasty registry* **73,** 548–554 (1995).

135. Villari, B. *et al.* Sex-dependent differences in left ventricular function and structure in chronic pressure overload. *Eur. Heart J.* **16,** 1410–1419 (1995).

136. Aurigemma, G. P., Silver, K. H., McLaughlin, M., Mauser, J. & Gaasch, W. H. Impact of chamber geometry and gender on left ventricular systolic function in patients > 60 years of age with aortic stenosis. *Am. J. Cardiol.* **74,** 794–798 (1994).

137. Cramariuc, D. *et al.* Sex differences in cardiovascular outcome during progression of aortic valve stenosis. *Heart* **101,** 209–214 (2015).

138. Nakamura, M. & Sadoshima, J. Mechanisms of physiological and pathological cardiac hypertrophy. *Nat. Rev. Cardiol.* **15,** 387–407 (2018).

139. Heineke, J. & Molkentin, J. D. Regulation of cardiac hypertrophy by intracellular signalling pathways. *Nat. Rev. Mol. Cell Biol.* **7,** 589–600 (2006).

140. Dworatzek, E., Baczko, I. & Kararigas, G. Effects of aging on cardiac extracellular matrix in men and women. *Proteomics Clin. Appl.* **10,** 84–91 (2016).

141. Gaignebet, L. & Kararigas, G. En route to precision medicine through the integration of biological sex into pharmacogenomics. *Clin Sci (Lond)* **131,** 329–342 (2017).

142. Chen, M. M., Lam, A., Abraham, J. A., Schreiner, G. F. & Joly, A. H. CTGF expression is induced by TGF- beta in cardiac fibroblasts and cardiac myocytes: a potential role in heart fibrosis. *J. Mol. Cell. Cardiol.* **32,** 1805–1819 (2000).

143. Chen, C.-C. & Lau, L. F. Functions and mechanisms of action of CCN matricellular proteins. *Int. J. Biochem. Cell Biol.* **41,** 771–783 (2009).

144. Daniels, A., van Bilsen, M., Goldschmeding, R., van der Vusse, G. J. & van Nieuwenhoven, F. A. Connective tissue growth factor and cardiac fibrosis. *Acta Physiol (Oxf)* **195,** 321–338 (2009).

145. Koentges, C. *et al.* Gene expression analysis to identify mechanisms underlying heart failure susceptibility in mice and humans. *Basic Res. Cardiol.* **113,** 8 (2018).

146. Dorn, L. E., Petrosino, J. M., Wright, P. & Accornero, F. CTGF/CCN2 is an autocrine regulator of cardiac fibrosis. *J. Mol. Cell. Cardiol.* **121,** 205–211 (2018).

147. Gordon, J. W., Shaw, J. A. & Kirshenbaum, L. A. Multiple facets of nf-kappab in the heart: To be or not to nf-kappab. *Circ Res* **108,** 1122–1132 (2011).

148. Saito, T. & And, G. A. C.-2. and nuclear factor-kappab in myocardium of end stage human heart failure. *Congest Heart Fail* **5,** 222–227 (1999).

149. Frantz, S. *et al.* Sustained activation of nuclear factor kappa B and activator protein 1 in chronic heart failure. *Cardiovasc. Res.* **57,** 749–756 (2003).

150. Sánchez-López, E. *et al.* CTGF promotes inflammatory cell infiltration of the renal interstitium by activating NF-kappaB. *J. Am. Soc. Nephrol.* **20,** 1513–1526 (2009).

151. Rodrigues-Diez, R. R. *et al.* The C-terminal module IV of connective tissue growth factor, through EGFR/Nox1 signaling, activates the NF-κB pathway and proinflammatory factors in vascular smooth muscle cells. *Antioxid. Redox Signal.* **22,** 29–47 (2015).

152. Chung, E. S., Packer, M., Lo, K. H., Fasanmade, A. A. & Randomized, W. J. double-blind, placebo-controlled, pilot trial of infliximab, a chimeric monoclonal antibody to tumor necrosis factor-alpha. *in patients with moderate-to-severe heart failure: Results of the anti-tnf therapy against congestive heart failure (attach) trial* **107,** 3133–3140 (2003).

153. Mann, D. L. *et al.* Targeted anticytokine therapy in patients with chronic heart failure: results of the Randomized Etanercept Worldwide Evaluation (RENEWAL). *Circulation* **109,** 1594–1602 (2004).

154. Beale, A. L., Meyer, P., Marwick, T. H., Lam, C. S. P. & Kaye, D. M. Sex differences in cardiovascular pathophysiology: why women are overrepresented in heart failure with preserved ejection fraction. *Circulation* **138,** 198–205 (2018).

155. Stolfo, D. *et al.* Sex-Based Differences in Heart Failure Across the Ejection Fraction Spectrum:

Phenotyping, and Prognostic and Therapeutic Implications. *JACC Heart Fail.* **7,** 505–515 (2019).

156. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).

157. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–574 (2013).

158. de Jonge, H. J. M. *et al.* Evidence based selection of housekeeping genes. *PLoS ONE* **2,** e898 (2007).

159. Lee, P. D., Sladek, R., Greenwood, C. M. T. & Hudson, T. J. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* **12,** 292–297 (2002).

160. Suzuki, T., Higgins, P. J. & Crawford, D. R. Control selection for RNA quantitation. *BioTechniques* **29,** 332–337 (2000).

161. Thellin, O. *et al.* Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* **75,** 291–295 (1999).

162. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

163. Lie, H. C. Towards breaking the curse of dimensionality in computational methods for the conformational analysis of molecules. *BMC Bioinformatics* **15,** (2014).

164. Antoniadis, A., Lambert-Lacroix, S. & Leblanc, F. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19,** 563–570 (2003).

165. Mak, M.-W. & Kung, S.-Y. in *Neural Information Processing* (eds. King, I., Wang, J., Chan, L.-W. & Wang, D.) **4232,** 314–323 (Springer Berlin Heidelberg, 2006).

166. MAQC Consortium *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161 (2006).

167. Shi, L. *et al.* Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6 Suppl 2,** S12 (2005).

168. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24,** 1162–1169 (2006).

169. Gopalappa, C., Das, T. K., Enkemann, S. & Eschrich, S. Removal of hybridization and scanning noise from microarrays. *IEEE Trans. Nanobioscience* **8,** 210–218 (2009).

170. Tu, Y., Stolovitzky, G. & Klein, U. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci USA* **99,** 14031–14036 (2002).

171. Balagurunathan, Y. *et al.* Noise factor analysis for cDNA microarrays. *J. Biomed. Opt.* **9,** 663–678 (2004).

172. Weng, L. *et al.* Rosetta error model for gene expression analysis. *Bioinformatics* **22,** 1111–1121

(2006).

173. Takeya, M. *et al.* Noise analysis of duplicated data on microarrays using mixture distribution modeling. *Opt. Rev.* **14,** 97–104 (2007).

174. Hong, H., Hong, Q., Liu, J., Tong, W. & Shi, L. Estimating relative noise to signal in DNA microarray data. *Int. J. Bioinform. Res. Appl.* **9,** 433–448 (2013).

175. Lee, Y. *et al.* Reducing system noise in copy number data using principal components of self-self hybridizations. *Proc Natl Acad Sci USA* **109,** E103-10 (2012).

176. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9,** e78644 (2014).

177. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4,** 14 (2009).

178. Bryant, P. A., Smyth, G. K., Robins-Browne, R. & Curtis, N. Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation. *PLoS ONE* **6,** e19556 (2011).

179. Friedman, D. B. Assessing signal-to-noise in quantitative proteomics: multivariate statistical analysis in DIGE experiments. *Methods Mol. Biol.* **854,** 31–45 (2012).

180. Yang, Y. *et al.* Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput. Struct. Biotechnol. J.* **9,** e201401002 (2014).

181. Shridhar, S. *et al.* Transcriptomic changes in CHO cells after adaptation to suspension growth in protein-free medium analysed by a species-specific microarray. *J. Biotechnol.* **257,** 13–21 (2017).

182. Graf, A., Dragosits, M., Gasser, B. & Mattanovich, D. Yeast systems biotechnology for the production of heterologous proteins. *FEMS Yeast Res.* **9,** 335–348 (2009).

183. Jayapal, K. P., Wlaschin, K. F., Hu, W. S. & Yap, M. G. S. Recombinant protein therapeutics from CHO Cells - 20 years and counting. *Chem* **103,** 40 (2007).

184. Walsh, G. Biopharmaceutical benchmarks 2010. *Nat. Biotechnol.* **28,** 917–924 (2010).

185. Brunner, D. *et al.* Serum-free cell culture: the serum-free media interactive online database. *ALTEX* **27,** 53–62 (2010).

186. Rodrigues, M. E. *et al.* Advances and Drawbacks of the Adaptation to Serum-Free Culture of CHO-K1 Cells for Monoclonal Antibody Production. *Appl. Biochem. Biotechnol.* **169,** 1279–1291 (2013).

187. Park, H., An, S. & Choe, T. Change of insulin-like growth factor gene expression in Chinese hamster ovary cells cultured in serum-free media. *Biotechnol. Bioprocess Eng.* **11,** 319–324 (2006).

188. Chong, W. P. K. *et al.* Metabolomics-driven approach for the improvement of Chinese hamster ovary cell growth: overexpression of malate dehydrogenase II. *J. Biotechnol.* **147,** 116–121 (2010).

189. Clarke, C. *et al.* Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics* **13,** 656 (2012).

190. Meleady, P. *et al.* Sustained productivity in recombinant Chinese hamster ovary (CHO) cell lines:

proteome analysis of the molecular basis for a process-related phenotype. *BMC Biotechnol.* **11,** 78 (2011).

191. Carlage, T. *et al.* Proteomic profiling of a high-producing Chinese hamster ovary cell culture. *Anal. Chem.* **81,** 7357–7362 (2009).

192. Kuystermans, D., Dunn, M. J. & Al-Rubeai, M. A proteomic study of cMyc improvement of CHO culture. *BMC Biotechnol.* **10,** 25 (2010).

193. Burleigh, S. C. *et al.* Synergizing metabolic flux analysis and nucleotide sugar metabolism to understand the control of glycosylation of recombinant protein in CHO cells. *BMC Biotechnol.* **11,** 95 (2011).

194. Shen, D. *et al.* Transcriptomic responses to sodium chloride-induced osmotic stress: a study of industrial fed-batch CHO cell cultures. *Biotechnol. Prog.* **26,** 1104–1115 (2010).

195. Kogenaru, S., Qing, Y., Guo, Y. & Wang, N. RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics* **13,** 629 (2012).

196. Matsumura, H. *et al.* High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS ONE* **5,** e12010 (2010).

197. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72 (2012).

198. Łabaj, P. P. *et al.* Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27,** i383-91 (2011).

199. Xu, W. *et al.* Human transcriptome array for high-throughput clinical studies. *Proc Natl Acad Sci USA* **108,** 3707–3712 (2011).

200. Zhang, Y., Akintola, O. S., Liu, K. J. A. & Sun, B. Membrane gene ontology bias in sequencing and microarray obtained by housekeeping-gene analysis. *Gene* **575,** 559–566 (2016).

201. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29,** 735–741 (2011).

202. Ernst, W. *et al.* Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. *Biotechnol. J.* **1,** 639–650 (2006).

203. Melville, M. *et al.* Development and characterization of a Chinese hamster ovary cell-specific oligonucleotide microarray. *Biotechnol. Lett.* **33,** 1773–1779 (2011).

204. Trummer, E. *et al.* Transcriptional profiling of phenotypically different Epo-Fc expressing CHO clones by cross-species microarray analysis. *Biotechnol. J.* **3,** 924–937 (2008).

205. Wlaschin, K. F. & Hu, W.-S. A scaffold for the Chinese hamster genome. *Biotechnol. Bioeng.* **98,** 429–439 (2007).

206. Baik, J. Y. *et al.* Initial transcriptome and proteome analyses of low culture temperature-induced expression in CHO cells producing erythropoietin. *Biotechnol. Bioeng.* **93,** 361–371 (2006).

207. De Leon Gatti, M., Wlaschin, K. F., Nissom, P. M., Yap, M. & Hu, W.-S. Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells

undergoing butyrate treatment. *J. Biosci. Bioeng.* **103,** 82–91 (2007).

208. Wlaschin, K. F. *et al.* EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol. Bioeng.* **91,** 592–606 (2005).

209. Wong, D. C. F. *et al.* Transcriptional profiling of apoptotic pathways in batch and fed-batch CHO cell cultures. *Biotechnol. Bioeng.* **94,** 373–382 (2006).

210. Vishwanathan, N. *et al.* Global insights into the Chinese hamster and CHO cell transcriptomes. *Biotechnol. Bioeng.* **112,** 965–976 (2015).

211. Rupp, O. *et al.* Construction of a public CHO cell line transcript database using versatile bioinformatics analysis pipelines. *PLoS ONE* **9,** e85568 (2014).

212. Becker, J. *et al.* Transcriptome analyses of CHO cells with the next-generation microarray CHO41K: development and validation by analysing the influence of the growth stimulating substance IGF-1 substitute LongR(3.). *J. Biotechnol.* **178,** 23–31 (2014).

213. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).

214. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47,** D330–D338 (2019).

215. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

216. Seth, G. *et al.* Large-scale gene expression analysis of cholesterol dependence in NS0 cells. *Biotechnol. Bioeng.* **90,** 552–567 (2005).

217. Gorfien, S. *et al.* Growth of NS0 cells in protein-free, chemically defined medium. *Biotechnol. Prog.* **16,** 682–687 (2000).

218. Sato, J. D. *et al.* Effects of proximate cholesterol precursors and steroid hormones on mouse myeloma growth in serum-free medium. Vitro Cell. *Dev. Biol* **24,** 1223 (1988).

219. Lee, T. C. *et al.* Elevation of glutathione levels and glutathione S-transferase activity in arsenic-resistant Chinese hamster ovary cells. *In Vitro Cell Dev Biol* **25,** 442–448 (1989).

220. Singhal, S. S. *et al.* Induction of glutathione S-transferase hGST 5.8 is an early response to oxidative stress in RPE cells. *Invest. Ophthalmol. Vis* **40,** 2652 (1999).

221. Pk, S. & Rg, M. Stress-induced proteins in immune response to cancer. *Curr. Top* **167,** 109 (1990).

222. Tezel, G. *et al.* Mechanisms of immune system activation in glaucoma: oxidative stress-stimulated antigen presentation by the retina and optic nerve head glia. *Invest. Ophthalmol. Vis. Sci.* **48,** 705–714 (2007).

223. Kensy, F. *et al.* Oxygen transfer phenomena in 48-well microtiter plates: determination by optical monitoring of sulfite oxidation and verification by real-time measurement during microbial growth. *Biotechnol. Bioeng.* **89,** 698–708 (2005).

224. Heinrich, C., Wolf, T., Kropp, C., Northoff, S. & Noll, T. Growth characterization of CHO DP-12 cell lines with different high passage histories. *BMC Proc.* **5 Suppl 8,** P29 (2011).

225. Ikonomou, L. & Kotton, D. N. Derivation of endodermal progenitors from pluripotent stem cells. *J. Cell. Physiol.* **230,** 246–258 (2015).

226. Green, M. D. *et al.* Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. *Nat. Biotechnol.* **29,** 267–272 (2011).

227. Huang, S. X. L. *et al.* Efficient generation of lung and airway epithelial cells from human pluripotent stem cells. *Nat. Biotechnol.* **32,** 84–91 (2014).

228. Kurmann, A. A. *et al.* Regeneration of thyroid function by transplantation of differentiated pluripotent stem cells. *Cell Stem Cell* **17,** 527–542 (2015).

229. Longmire, T. A. *et al.* Efficient derivation of purified lung and thyroid progenitors from embryonic stem cells. *Cell Stem Cell* **10,** 398–411 (2012).

230. Mou, H. *et al.* Generation of multipotent lung and airway progenitors from mouse ESCs and patient-specific cystic fibrosis iPSCs. *Cell Stem Cell* **10,** 385–397 (2012).

231. Parent, A. V. *et al.* Generation of functional thymic epithelium from human embryonic stem cells that supports host T cell development. *Cell Stem Cell* **13,** 219–229 (2013).

232. Sun, X. *et al.* Directed differentiation of human embryonic stem cells into thymic epithelial progenitor-like cells reconstitutes the thymic microenvironment in vivo. *Cell Stem Cell* **13,** 230–236 (2013).

233. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).

234. Antonica, F. *et al.* Generation of functional thyroid from embryonic stem cells. *Nature* **491,** 66–71 (2012).

235. Kubo, A. *et al.* Pdx1 and Ngn3 overexpression enhances pancreatic differentiation of mouse ES cell-derived endoderm population. *PLoS ONE* **6,** e24058 (2011).

236. Kyba, M., Perlingeiro, R. C. R. & Daley, G. Q. HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic progenitors. *Cell* **109,** 29–37 (2002).

237. Petros, T. J., Maurer, C. W. & Anderson, S. A. Enhanced derivation of mouse ESC-derived cortical interneurons by expression of Nkx2.1. *Stem Cell Res.* **11,** 647–656 (2013).

238. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Comput. Biol.* **10,** e1003734 (2014).

239. Pusuluri, S. T., Lang, A. H., Mehta, P. & Castillo, H. E. *Cellular reprogramming dynamics follow a simple one-dimensional reaction coordinate. arXiv 1505.03889 [q-bio.* (MN], 2015).

240. Goss, A. M. *et al.* Wnt2/2b and beta-catenin signaling are necessary and sufficient to specify lung progenitors in the foregut. *Dev. Cell* **17,** 290–298 (2009).

241. Harris-Johnson, K. S., Domyan, E. T., Vezina, C. M. & Sun, X. beta-Catenin promotes respiratory progenitor identity in mouse foregut. *Proc Natl Acad Sci USA* **106,** 16287–16292 (2009).

242. Millien, G. *et al.* Characterization of the mid-foregut transcriptome identifies genes regulated during lung bud induction. *Gene Expr. Patterns* **8,** 124–139 (2008).

243. Westerlund, J. *et al.* Expression of Islet1 in thyroid development related to budding, migration, and fusion of primordia. *Dev. Dyn.* **237,** 3820–3829 (2008).

244. Ishii, J. *et al.* PROX1 promotes secretory granule formation in medullary thyroid cancer cells. *Endocrinology* **157,** 1289–1298 (2016).

245. Loh, K. M. *et al.* Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* **14,** 237–252 (2014).

246. Perlman, R. L. Mouse models of human disease: An evolutionary perspective. *Evol. Med. Public Health* **2016,** 170–176 (2016).

247. Fields, S. & Johnston, M. Cell biology. Whither model organism research? *Science* **307,** 1885–1886 (2005).

248. Pinnapureddy, A. R. *et al.* Large animal models of rare genetic disorders: sheep as phenotypically relevant models of human genetic disease. *Orphanet J. Rare Dis.* **10,** 107 (2015).

249. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132,** 661–680 (2008).

250. Domyan, E. T. *et al.* Signaling through BMP receptors promotes respiratory identity in the foregut via repression of Sox2. *Development* **138,** 971–981 (2011).

251. Serra, M. *et al.* Pluripotent stem cell differentiation reveals distinct developmental pathways regulating lung- versus thyroid-lineage specification. *Development* **144,** 3879–3893 (2017).

252. Firth, A. L. *et al.* Generation of multiciliated cells in functional airway epithelia from human induced pluripotent stem cells. *Proc Natl Acad Sci U S A* **111,** E1723–E1730 (2014).

253. Gotoh, S. *et al.* Generation of alveolar epithelial spheroids via isolated progenitor cells from human pluripotent stem cells. *Stem Cell Reports* **3,** 394–403 (2014).

254. Konishi, S. *et al.* Directed Induction of Functional Multi-ciliated Cells in Proximal Airway Epithelial Spheroids from Human Pluripotent Stem Cells. *Stem Cell Reports* **6,** 18–25 (2016).

255. Jacob, A. *et al.* Differentiation of Human Pluripotent Stem Cells into Functional Lung Alveolar Epithelial Cells. *Cell Stem Cell* **21,** 472-488.e10 (2017).

256. McCauley, K. B. *et al.* Efficient Derivation of Functional Human Airway Epithelium from Pluripotent Stem Cells via Temporal Regulation of Wnt Signaling. *Cell Stem Cell* **20,** 844-857.e6 (2017).

257. Tsao, P.-N. *et al.* Notch signaling controls the balance of ciliated and secretory cell fates in developing airways. *Development* **136,** 2297–2307 (2009).

258. Hashimoto, S. *et al.* beta-Catenin-SOX2 signaling regulates the fate of developing airway epithelium. *J Cell Sci* **125,** 932–942 (2012).

259. Mucenski, M. L. *et al.* beta-Catenin is required for specification of proximal/distal cell fate during lung morphogenesis. *J. Biol. Chem.* **278,** 40231–40238 (2003).

260. Shu, W. *et al.* Wnt/beta-catenin signaling acts upstream of N-myc, BMP4, and FGF signaling to regulate proximal-distal patterning in the lung. *Dev. Biol.* **283,** 226–239 (2005).

261. Desai, T. J., Malpel, S., Flentke, G. R., Smith, S. M. & Cardoso, W. V. Retinoic acid selectively regulates Fgf10 expression and maintains cell identity in the prospective lung field of the developing foregut. *Dev. Biol.* **273,** 402–415 (2004).

262. Lazzaro, D., Price, M., Defelice, M. & Dilauro, R. The Transcription Factor-Ttf-1 Is Expressed at the Onset of Thyroid and Lung Morphogenesis and in Restricted Regions of the Fetal Brain. *Development* **113,** 1093–1104 (1991).

263. Hawkins, F. *et al.* Prospective isolation of NKX2-1-expressing human lung progenitors derived from pluripotent stem cells. *J. Clin. Invest.* **127,** 2277–2294 (2017).

264. Fagman, H. *et al.* Gene expression profiling at early organogenesis reveals both common and diverse mechanisms in foregut patterning. *Dev. Biol.* **359,** 163–175 (2011).

265. Ikonomou, L., Wagner, D. E., Turner, L. & Weiss, D. J. Translating Basic Research into Safe and Effective Cell-based Treatments for Respiratory Diseases. *Ann Am Thoracic Society* **16,** 657–668 (2019).

266. Dye, B. R. *et al.* In vitro generation of human pluripotent stem cell derived lung organoids. *elife* **4,** (2015).

267. Rankin, S. A. *et al.* A Retinoic Acid-Hedgehog Cascade Coordinates Mesoderm-Inducing Signals and Endoderm Competence during Lung Specification. *Cell Rep.* **16,** 66–78 (2016).

268. Rankin, S. A. *et al.* Timing is everything: Reiterative Wnt, BMP and RA signaling regulate developmental competence during endoderm organogenesis. *Dev. Biol.* **434,** 121–132 (2018).

269. Mahoney, J. E., Mori, M., Szymaniak, A. D., Varelas, X. & Cardoso, W. V. The hippo pathway effector Yap controls patterning and differentiation of airway epithelial progenitors. *Dev. Cell* **30,** 137–150 (2014).

270. Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158,** 903–915 (2014).

271. D'Alessio, A. C. *et al.* A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem Cell Reports* **5,** 763–775 (2015).

272. Morris, S. A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158,** 889–902 (2014).

273. Pusuluri, S. T., Lang, A. H., Mehta, P. & Castillo, H. E. Cellular reprogramming dynamics follow a simple 1D reaction coordinate. *Phys Biol* **15,** 016001 (2018).

274. Boring, M. A. *et al.* Prevalence of Arthritis and Arthritis-Attributable Activity Limitation by Urban-Rural County Classification - United States, 2015. *MMWR Morb Mortal Wkly Rep* **66,** 527–532 (2017).

275. Johnson, V. L. & Hunter, D. J. The epidemiology of osteoarthritis. *Best Pract. Res. Clin. Rheumatol.* **28,** 5–15 (2014).

276. Lopez, A. D. & Murray, C. C. J. L. The global burden of disease, 1990–2020. *Nat. Med.* **4,** 1241–

1243 (1998).

277. Woolf, A. D. & Pfleger, B. Burden of major musculoskeletal conditions. *Bull. World Health Organ.* **81,** 646–656 (2003).

278. Rabenda, V. *et al.* Prevalence and impact of osteoarthritis and osteoporosis on health-related quality of life among active subjects. *Aging Clin Exp Res* **19,** 55 (2007).

279. Yelin, E. *et al.* Medical care expenditures and earnings losses among persons with arthritis and other rheumatic conditions in 2003, and comparisons with 1997. *Arthritis Rheum.* **56,** 1397–1407 (2007).

280. Goldring, M. B. *et al.* Roles of inflammatory and anabolic cytokines in cartilage metabolism: signals and multiple effectors converge upon MMP-13 regulation in osteoarthritis. *Eur. Cell. Mater.* **21,** 202–220 (2011).

281. Pap, T. & Korb-Pap, A. Cartilage damage in osteoarthritis and rheumatoid arthritis--two unequal siblings. *Nat. Rev. Rheumatol.* **11,** 606–615 (2015).

282. al-Qattan, M. M., Posnick, J. C., Lin, K. Y. & Thorner, P. Fetal tendon healing: development of an experimental model. *Plast. Reconstr. Surg.* **92,** 1155–60; discussion 1161 (1993).

283. Degen, K. E. & Gourdie, R. G. Embryonic wound healing: a primer for engineering novel therapies for tissue repair. *Birth Defects Res. C Embryo Today* **96,** 258–270 (2012).

284. Kumahashi, N. *et al.* Involvement of ATP, increase of intracellular calcium and the early expression of c-fos in the repair of rat fetal articular cartilage. *Cell Tissue Res.* **317,** 117–128 (2004).

285. Longaker, M. T. *et al.* Fetal fracture healing in a lamb model. *Plast. Reconstr. Surg.* **90,** 161–71; discussion 172 (1992).

286. Longaker, M. T. *et al.* Studies in fetal wound healing VI. *Second and early third trimester fetal wounds demonstrate rapid collagen deposition without scar formation* **25,** 63 (1990).

287. Namba, R. S., Meuli, M., Sullivan, K. M., Le, A. X. & Adzick, N. S. Spontaneous repair of superficial defects in articular cartilage in a fetal lamb model. *J. Bone Joint Surg. Am.* **80,** 4–10 (1998).

288. Stone, C. A. Unravelling the secrets of foetal wound healing: an insight into fracture repair in the mouse foetus and perspectives for clinical application. *Br J Plast Surg* **53,** 337 (2000).

289. Wagner, W., Reichl, J., Wehrmann, M. & Zenner, H. P. Neonatal rat cartilage has the capacity for tissue regeneration. *Wound Repair Regen.* **9,** 531–536 (2001).

290. Walker, E. A., Verner, A., Flannery, C. R. & Archer, C. W. Cellular responses of embryonic hyaline cartilage to experimental wounding in vitro. *Journal of orthopaedic research* **18,** 25 (2000).

291. Decker, R. S. *et al.* Cell origin, volume and arrangement are drivers of articular cartilage formation, morphogenesis and response to injury in mouse limbs. *Dev. Biol.* **426,** 56–68 (2017).

292. Jenner, F. *et al.* Differential gene expression of the intermediate and outer interzone layers of developing articular cartilage in murine embryos. *Stem Cells Dev.* **23,** 1883–1898 (2014).

293. Gerjo JVM van Osch, F. J. Laser capture microdissection of murine interzone cells: layer selection

and prediction of RNA yield. *J. Stem Cell Res. Ther.* **04,** (2014).

294. Lo, D. D., Zimmermann, A. S., Nauta, A., Longaker, M. T. & Lorenz, H. P. Scarless fetal skin wound healing update. *Birth Defect Res C* **96,** 237 (2012).

295. Cowin, A. J., Brosnan, M. P., Holmes, T. M. & Ferguson, M. W. Endogenous inflammatory response to dermal wound healing in the fetal and adult mouse. *Dev. Dyn.* **212,** 385–393 (1998).

296. Ferguson, M. W. J. & O'Kane, S. Scar-free healing: from embryonic mechanisms to adult therapeutic intervention. *Philos. Trans. R. Soc. Lond. B. Biol. Sci* **359,** 839–850 (2004).

297. Almeida-Porada, G., Porada, C. & Zanjani, E. D. Plasticity of human stem cells in the fetal sheep model of human stem cell transplantation. *Int. J. Hematol.* **79,** 1–6 (2004).

298. Jeanblanc, C. *et al.* Temporal definition of haematopoietic stem cell niches in a large animal model of in utero stem cell transplantation. *Br. J. Haematol.* **166,** 268–278 (2014).

299. Dorotka, R., Bindreiter, U., Macfelda, K., Windberger, U. & Nehrer, S. Marrow stimulation and chondrocyte transplantation using a collagen matrix for cartilage repair. *Osteoarthr. Cartil.* **13,** 655–664 (2005).

300. Entrican, G., Wattegedera, S. R. & Griffiths, D. J. Exploiting ovine immunology to improve the relevance of biomedical models. *Mol. Immunol.* **66,** 68–77 (2015).

301. Mrugala, D. *et al.* Phenotypic and functional characterisation of ovine mesenchymal stem cells: application to a cartilage defect model. *Ann. Rheum. Dis.* **67,** 288–295 (2008).

302. van Turnhout, M. C. *et al.* Postnatal development of depth-dependent collagen density in ovine articular cartilage. *BMC Dev. Biol.* **10,** 108 (2010).

303. van Turnhout, M. C. *et al.* Postnatal development of collagen structure in ovine articular cartilage. *BMC Dev. Biol.* **10,** 62 (2010).

304. Lu, Y., Markel, M. D., Swain, C. & Kaplan, L. D. Development of partial thickness articular cartilage injury in an ovine model. *J. Orthop. Res* **24,** 1974 (2006).

305. Munirah, S. *et al.* Articular cartilage restoration in load-bearing osteochondral defects by implantation of autologous chondrocyte-fibrin constructs: an experimental study in sheep. *J. Bone Joint Surg. Br.* **89,** 1099–1109 (2007).

306. Xue, X., Zheng, Q., Wu, H., Zou, L. & Li, P. Different responses to mechanical injury in neonatal and adult ovine articular cartilage. *Biomed Eng Online* **12,** 53 (2013).

307. Almeida-Porada, G. *et al.* The human-sheep chimeras as a model for human stem cell mobilization and evaluation of hematopoietic grafts' potential. *Exp. Hematol.* **35,** 1594–1600 (2007).

308. Kim, J., Zanjani, E. D., Jeanblanc, C. M., Goodrich, A. D. & Hematti, P. Generation of CD34+ cells from human embryonic stem cells using a clinically applicable methodology and engraftment in the fetal sheep model. *Exp. Hematol.* **41,** 749-758.e5 (2013).

309. Kuypers, E. *et al.* White matter injury following fetal inflammatory response syndrome induced by chorioamnionitis and fetal sepsis: lessons from experimental ovine models. *Early Hum. Dev.* **88,** 931–936 (2012).

310. Liechty, K. W. *et al.* Human mesenchymal stem cells engraft and demonstrate site-specific differentiation after in utero transplantation in sheep. *Nat. Med.* **6,** 1282–1286 (2000).

311. Porada, C. D. *et al.* Gestational age of recipient determines pattern and level of transgene expression following in utero retroviral gene transfer. *Mol. Ther.* **11,** 284–293 (2005).

312. Bruns, J., Kampen, J., Kahrs, J. & Plitz, W. Achilles tendon rupture: experimental results on spontaneous repair in a sheep-model. *Knee surgery, sports traumatology, arthroscopy* **8,** 364 (2000).

313. Russo, V. *et al.* Cellular and molecular maturation in fetal and adult ovine calcaneal tendons. *J. Anat.* **226,** 126–142 (2015).

314. Maddox, J. F., Mackay, C. R. & Brandon, M. R. Ontogeny of ovine lymphocytes. *An immunohistological study on the development of T lymphocytes in the sheep embryo and fetal thymus* **62,** 97 (1987).

315. Miyasaka, M. & Trnka, Z. Lymphocyte Migration and Differentiation in a Large? *nimal Model: The Sheep* **91,** 87 (1986).

316. Osburn, B. I. in *Adv* (Exp. Med. Biol. 137, 1981).

317. Sawyer, M., Moe, J. & Osburn, B. I. Ontogeny of immunity and leukocytes in the ovine fetus and elevation of immunoglobulins related to congenital infection. *American Journal of Veterinary Research* **39,** 643 (1978).

318. Coordinators, N. R. NCBI Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **46,** (2018).

319. Stenberg, J., Rüetschi, U., Skiöldebrand, E., Kärrholm, J. & Lindahl, A. Quantitative proteomics reveals regulatory differences in the chondrocyte secretome from human medial and lateral femoral condyles in osteoarthritic patients. *Proteome Sci.* **11,** 43 (2013).

320. Wilson, R. & Bateman, J. F. Cartilage proteomics: Challenges, solutions and recent advances. *Prot* **2,** 251 (2008).

321. Ritter, S. Y. *et al.* Proteomic analysis of synovial fluid from the osteoarthritic knee: comparison with transcriptome analyses of joint tissues. *Arthritis Rheum.* **65,** 981–992 (2013).

322. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

323. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** Article3 (2004).

324. Sun, W. *et al.* PPM1A and PPM1B act as IKK? phosphatases to terminate TNF?-induced IKK?-NF-? *activation. Cellular Signalling* **21,** 95 (2009).

325. Murdoch, D. J. & Chow, E. D. A graphical display of large correlation matrices. *The American Statistician* **50,** 178–180 (1996).

326. Beier, F. & Loeser, R. F. Biology and pathology of Rho GTPase, PI-3 kinase-Akt, and MAP kinase signaling pathways in chondrocytes. *J. Cell. Biochem.* **110,** 573–580 (2010).

327. Wang, J. *et al.* The self-limiting dynamics of TGF-β signaling in silico and in vitro, with negative feedback through PPM1A upregulation. *PLoS Comput. Biol.* **10,** e1003573 (2014).

328. Duan, X., Liang, Y.-Y., Feng, X.-H. & Lin, X. Protein serine/threonine phosphatase PPM1A dephosphorylates Smad1 in the bone morphogenetic protein signaling pathway. *J. Biol. Chem.* **281,** 36526–36532 (2006).

329. Wang, J. R. *et al.* Signaling Cascades Governing Cdc42-Mediated Chondrogenic Differentiation and Mensenchymal Condensation. *Genetics* **202,** 1055–1069 (2016).

330. Rohani, M. G., Pilcher, B. K., Chen, P. & Parks, W. C. Cdc42 inhibits ERK-mediated collagenase-1 (MMP-1) expression in collagen-activated human keratinocytes. *J. Invest. Dermatol.* **134,** 1230–1237 (2014).

331. Benink, H. A. & Bement, W. M. Concentric zones of active RhoA and Cdc42 around single cell wounds. *J. Cell Biol.* **168,** 429–439 (2005).

332. Liu-Bryan, R. & Terkeltaub, R. Emerging regulators of the inflammatory process in osteoarthritis. *Nat. Rev. Rheumatol.* **11,** 35–44 (2015).

333. Nefla, M., Holzinger, D., Berenbaum, F. & Jacques, C. The danger from within: alarmins in arthritis. *Nat. Rev. Rheumatol.* **12,** 669–683 (2016).

334. van den Bosch, M. H. *et al.* Induction of canonical wnt signaling by the alarmins S100A8/A9 in murine knee joints: implications for osteoarthritis. *Arthritis Rheumatol.* **68,** 152–163 (2016).

335. Schelbergen, R. F. P. *et al.* Alarmins S100A8 and S100A9 elicit a catabolic effect in human osteoarthritic chondrocytes that is dependent on Toll-like receptor 4. *Arthritis Rheum.* **64,** 1477–1487 (2012).

336. Nakashima, M. *et al.* Role of S100A12 in the pathogenesis of osteoarthritis. *Biochem. Biophys. Res. Commun.* **422,** 508–514 (2012).

337. Dunkel, Y. *et al.* STAT3 protein up-regulates G?-interacting vesicle-associated protein (GIV)/Girdin expression, and GIV enhances STAT3 activation in a positive feedback loop during wound healing and tumor invasion/metastasis. *Journal of Biological Chemistry* **287,** 41667 (2012).

338. Ghosh, P., Garcia-Marcos, M., Bornheimer, S. J. & Farquhar, M. G. Activation of Galphai3 triggers cell migration via regulation of GIV. *J. Cell Biol.* **182,** 381–393 (2008).

339. Kennedy, J. M. *et al.* CCDC88B is a novel regulator of maturation and effector functions of T cells during pathological inflammation. *J. Exp. Med.* **211,** 2519–2535 (2014).

340. Lopez-Sanchez, I. *et al.* GIV/Girdin is a central hub for profibrogenic signalling networks during liver fibrosis. *Nat. Commun.* **5,** 4451 (2014).

341. Chen, J., Crawford, R. & Xiao, Y. Vertical inhibition of the PI3K/Akt/mTOR pathway for the treatment of osteoarthritis. *J. Cell. Biochem.* **114,** 245–249 (2013).

342. Fujita, T. *et al.* Runx2 induces osteoblast and chondrocyte differentiation and enhances their migration by coupling with PI3K-Akt signaling. *J. Cell Biol.* **166,** 85–95 (2004).

343. Greene, M. A. & Loeser, R. F. Function of the chondrocyte PI-3 kinase-Akt signaling pathway is stimulus dependent. *Osteoarthr. Cartil.* **23,** 949–956 (2015).

344. Kita, K., Kimura, T., Nakamura, N., Yoshikawa, H. & Nakano, T. PI3K/Akt signaling as a key regulatory pathway for chondrocyte terminal differentiation. *Genes Cells* **13,** 839–850 (2008).

345. Litherland, G. J. *et al.* Synergistic collagenase expression and cartilage collagenolysis are phosphatidylinositol 3-kinase/Akt signaling-dependent. *J. Biol. Chem.* **283,** 14221–14229 (2008).

346. Starkman, B. G., Cravero, J. D., Delcarlo, M. & Loeser, R. F. IGF-I stimulation of proteoglycan synthesis by chondrocytes requires activation of the PI 3-kinase pathway but not ERK MAPK. *Biochem. J.* **389,** 723–729 (2005).

347. Xu, J., Yi, Y., Li, L., Zhang, W. & Wang, J. Osteopontin induces vascular endothelial growth factor expression in articular cartilage through PI3K/AKT and ERK1/2 signaling. *Mol. Med. Report.* **12,** 4708–4712 (2015).

348. Kozhemyakina, E. *et al.* Identification of a Prg4-expressing articular cartilage progenitor cell population in mice. *Arthritis Rheumatol.* **67,** 1261–1273 (2015).

349. Rosmarin, A. G., Resendes, K. K., Yang, Z., McMillan, J. N. & Fleming, S. L. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol. Dis.* **32,** 143–154 (2004).

350. Juan, W. C. & Hong, W. Targeting the hippo signaling pathway for tissue regeneration and cancer therapy. *Genes (Basel)* **7,** (2016).

351. Ueda, A., Akagi, T. & Yokota, T. GA-Binding Protein Alpha Is Involved in the Survival of Mouse Embryonic Stem Cells. *Stem Cells* **35,** 2229–2238 (2017).

352. Bobick, B. E., Matsche, A. I., Chen, F. H. & Tuan, R. S. *The ERK5 and ERK1/2 signaling pathways play opposing regulatory roles during chondrogenesis of adult human bone marrow-derived multipotent progenitor cells*. (J. Cell. Physiol. n/a–n/a, 2010).

353. Emmert, M. Y. *et al.* Intramyocardial transplantation and tracking of human mesenchymal stem cells in a novel intra-uterine pre-immune fetal sheep myocardial infarction model: a proof of concept study. *PLoS ONE* **8,** e57759 (2013).

354. Mackay, C. R., Maddox, J. F. & Brandon, M. R. Thymocyte subpopulations during early fetal development in sheep. *J. Immunol.* **136,** 1592–1599 (1986).

355. Silverstein, A. M., Uhr, J. W., Kraner, K. & Lukes, R. in *J* (Exp. Med. 117, 1963).

356. Silverstein, A. M., Prendergast, R. & Kraner, K. in *of skin homografts by the fetal lamb* (ed. Rejection, I. V.) (J. Exp. Med. 119, 1964).

357. Kumta, S. *et al.* Acute inflammation in foetal and adult sheep: the response to subcutaneous injection of turpentine and carrageenan. *Br. J. Plast. Surg.* **47,** 360–368 (1994).

358. Moss, T. J. M. *et al.* Experimental amniotic fluid infection in sheep: effects of Ureaplasma parvum serovars 3 and 6 on preterm or term fetal sheep. *Am. J. Obstet. Gynecol.* **198,** 122.e1–8 (2008).

359. Nitsos, I. *et al.* Chronic exposure to intra-amniotic lipopolysaccharide affects the ovine fetal brain. *J. Soc. Gynecol. Investig.* **13,** 239–247 (2006).

360. Dziegielewska, K. M. *et al.* Acute-phase cytokines IL-1beta and TNF-alpha in brain development. *Cell Tissue Res.* **299,** 335–345 (2000).

361. Dougherty, E. R. & Shmulevich, I. On the limitations of biological knowledge. *Curr. Genomics* **13,** 574–587 (2012).

362. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry* **80,** 552–561 (2016).

363. Patrick, E. *et al.* A multi-step classifier addressing cohort heterogeneity improves performance of prognostic biomarkers in three cancer types. *Oncotarget* **8,** 2807–2815 (2017).

364. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

365. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* **100,** 8418–8423 (2003).

366. Gyanchandani, R. *et al.* Intratumor heterogeneity affects gene expression profile test prognostic risk stratification in early breast cancer. *Clin. Cancer Res.* **22,** 5362–5369 (2016).

367. Stigliani, S. *et al.* High genomic instability predicts survival in metastatic high-risk neuroblastoma. *Neoplasia* **14,** 823–832 (2012).

368. Kocak, H. *et al.* Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma. *Cell Death Dis.* **4,** e586 (2013).

369. Theissen, J. *et al.* Chromosome 17/17q gain and unaltered profiles in high resolution array-CGH are prognostically informative in neuroblastoma. *Genes Chromosomes Cancer* **53,** 639–649 (2014).

370. Coco, S. *et al.* Age-dependent accumulation of genomic aberrations and deregulation of cell cycle and telomerase genes in metastatic neuroblastoma. *Int. J. Cancer* **131,** 1591–1600 (2012).

371. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* **16,** 133 (2015).

372. Kivelä, M. *et al.* Multilayer networks. *J. Complex Netw.* **2,** 203–271 (2014).

373. Léger, J.-B. Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates. (2016).

374. Côme, E. & Latouche, P. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling* **15,** 564–589 (2015).

375. Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22,** 719–725 (2000).

376. Meilă, M. in *Learning Theory and Kernel Machines* (eds. Schölkopf, B. & Warmuth, M. K.) **2777,** 173–187 (Springer Berlin Heidelberg, 2003).

377. Meilă, M. Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98,** 873–895 (2007).

378. Heinze, G. & Dunkler, D. Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Stat. Med.* **27,** 6455–6469 (2008).

379. Johnson, K. & Lin, S. Call to work together on microarray data analysis. *Nature* **411,** 885 (2001).

380. Tilstone, C. DNA microarrays: vital statistics. *Nature* **424,** 610–612 (2003).

381. Going for algorithm gold. *Nat. Methods* **5,** 659–659 (2008).

382. McGuire, S. World cancer report 2014. geneva, switzerland: world health organization, international agency for research on cancer, WHO press, 2015. *Adv. Nutr.* **7,** 418–419 (2016).

383. Maris, J. M., Hogarty, M. D., Bagatell, R. & Cohn, S. L. Neuroblastoma. *Lancet* **369,** 2106–2120 (2007).

384. Pfitzner, D., Leibbrandt, R. & Powers, D. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowl. Inf. Syst.* **19,** 361–394 (2009).

385. Baali, I., Acar, D. A. E., Aderinwale, T. W., HafezQorani, S. & Kazan, H. Predicting clinical outcomes in neuroblastoma with genomic data integration. *Biol. Direct* **13,** 20 (2018).

386. Suo, C. *et al.* Accumulation of potential driver genes with genomic alterations predicts survival of high-risk neuroblastoma patients. *Biol. Direct* **13,** 14 (2018).

387. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).

388. Zuo, S., Dai, G. & Ren, X. Identification of a 6-gene signature predicting prognosis for colorectal cancer. *Cancer Cell Int.* **19,** 6 (2019).

389. Raue, F. & Frank-Raue, K. Thyroid Cancer: Risk-Stratified Management and Individualized Therapy. *Clin. Cancer Res.* **22,** 5012–5021 (2016).

390. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159,** 676–690 (2014).

391. Chi, J.-T. *et al.* Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med.* **3,** e47 (2006).

392. Fardin, P. *et al.* The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. *BMC Genomics* **10,** 474 (2009).

393. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **46,** D649–D655 (2018).

394. Eblen, J. D. *et al.* Graph Algorithms for Integrated Biological Analysis, with Applications to Type 1 Diabetes Data. 207–222

395. Jiang, H. *et al.* Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* **5,** 81 (2004).

396. Claus, B. L. & Underwood, D. J. Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* **7,** 957–966 (2002).

397. Augen, J. The evolving role of information technology in the drug discovery process. *Drug Discov. Today* **7,** 315–323 (2002).

398. Dimitrieva, S., Schlapbach, R. & Rehrauer, H. Prognostic value of cross-omics screening for kidney clear cell renal cancer survival. *Biol. Direct* **11,** 68 (2016).

399. Dai, L., Gao, X., Guo, Y., Xiao, J. & Zhang, Z. Bioinformatics clouds for big data manipulation. *Biol. Direct* **7,** 43; discussion 43 (2012).

400. Tarczy-Hornoch, P. *et al.* Meeting clinician information needs by integrating access to the medical record and knowledge resources via the Web. *Proc. AMIA Annu. Fall Symp.* 809–813 (1997).

401. Francescatto, M. *et al.* Multi-omics integration for neuroblastoma clinical endpoint prediction. *Biol. Direct* **13,** 5 (2018).

402. Tranchevent, L.-C. *et al.* Predicting clinical outcome of neuroblastoma patients using an integrative network-based approach. *Biol. Direct* **13,** 12 (2018).

403. Catarci, T. & Lenzerini, M. REPRESENTING AND USING INTERSCHEMA KNOWLEDGE IN COOPERATIVE INFORMATION SYSTEMS. *International Journal of Cooperative Information Systems* **02,** 375–398 (1993).

404. Rahm, E. & Bernstein, P. A. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal* **10,** 334–350 (2001).

405. Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C. & Wallach, D. A. Bigtable: A Distributed Storage System for Structured Data. *ACM Trans. Comput. Syst.* **26,** 4:1-4:26 (2008).

406. Curé, O., Hecht, R., Le Duc, C. & Lamolle, M. in *Database and expert systems applications* (eds. Hameurlain, A., Liddle, S. W., Schewe, K.-D. & Zhou, X.) **6860,** 481–495 (Springer Berlin Heidelberg, 2011).

407. Meeker, W. Q. & Escobar, L. A. Statistical Methods for Reliability Data. 26–45 (1998).

408. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

409. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).

410. Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res. Treat.* **22,** 207–219 (1992).

411. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2,** 59–77 (2007).

412. Simes, R. J. Treatment selection for cancer patients: Application of statistical decision theory to the treatment of advanced ovarian cancer. *Journal of Chronic Diseases* **38,** 171–186 (1985).

413. Zhang, H. *et al.* Data Integration through Ontology-Based Data Access to Support Integrative Data Analysis: A Case Study of Cancer Survival. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2017,** 1300–1303 (IEEE, 2017).

414. Vijayarani, D. S. & Dhayanand, M. S. Kidney Disease Prediction Using Svm and Ann Algorithms. (2015).

415. Gupta, S. *et al.* Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open* **4,** e004007 (2014).

416. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32,** 926–932 (2014).

417. Zang, C. *et al.* High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* **7,** 11305 (2016).

418. Tarca, A. L., Bhatti, G. & Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE* **8,** e79217 (2013).

419. Bayerlová, M. *et al.* Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics* **16,** 334 (2015).

420. Zhu, B. *et al.* Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. *Sci. Rep.* **7,** 16954 (2017).