STATegra: a comprehensive multi-omics dataset of B-cell differentiation in mouse

David Gomez-Cabrero[1,2,3,a], Sonia Tarazona[4,a], Isabel Ferreirós-Vidal[5,a], Ricardo N. Ramirez[6,a], Carlos Company[7,a], Andreas Schmidt[8,a], Theo Reijmers[9,a], Veronica von Saint Paul[10,a], Franscesco Marabita[2], Javier Rodríguez-Ubreva[7], Antonio Garcia-Gomez[7], Thomas Carroll[5], Lee Cooper[5], Ziwei Liang[5], Gopuraja Dharmalingam[5], Leandro Balzano-Nogueira[11], Vicenzo Lagani[12,13], Ioannis Tsamardinos[12,13,b], Michael Lappe[14,b], Dieter Maier[10,b], Johan A. Westerhuis[15,16,b], Thomas Hankemeier[9,b], Axel Imhof[8,b], Esteban Ballestar[7,b], Ali Mortazavi[6,b], Matthias Merkenschlager[5,b], Jesper Tegner[2,3,17,b], Ana Conesa[11,b]

[1]Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA. Pamplona, Spain
[2]Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden
[3]Science for Life Laboratory, Solna, Sweden
[4]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain
[5]MRC London Institute of Medical Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, UK
[6]Department of Developmental and Cell Biology and Center for Complex Biological Systems, University of California, Irvine, CA, USA
[7]Chromatin and Disease Group, Cancer Epigenetics and Biology Programme (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[8] Protein Analysis Unit, Biomedical Center, Ludwig Maximilian University of Munich, Munich, Germany
[9] Division of Analytical Biosciences, Leiden Academic Center for Drug Research, Leiden University, Leiden, The Netherlands
[10] Biomax Informatics AG, Planegg, Germany
[11]Microbiology and Cell Science Department, Genetics Institute, University of Florida, Gainesville, Florida, USA
[12]Computer Science Department, University of Crete, Heraklion, Greece
[13]Gnosis Data Analysis PC, Heraklion, Greece
[14]QIAGEN Aarhus A/S, Silkeborgvej 2, 8000 Aarhus, Denmark
[15]Centre for Human Metabolomics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Potchefstroom, South Africa
[16]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
[17]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Saudi Arabia.
[a] These authors contributed equally to this work
[b] These authors jointly supervised this work

Contact: Ana Conesa: aconesa@ufl.edu; Jesper Tegner: jesper.tegner@kaust.edu.sa; Matthias Merlenschlager: matthias.merkenschlager@lms.mrc.ac.uk

Abstract

Multi-omics approaches use a diversity of high-throughput technologies to profile the different molecular layers of living cells. Ideally, the integration of this information should result in comprehensive systems models of cellular physiology and regulation. However, most multiomics projects still include a limited number of molecular assays and there have been very few multi-omic studies that evaluate dynamic processes such as cellular growth, development and adaptation. Hence, we lack formal analysis methods and comprehensive multi-omics datasets that can be leveraged to develop true multi-layered models for dynamic cellular systems. Here we present the STATegra multi-omics dataset that combines measurements from up to 10 different omics technologies applied to the same biological system, namely the well-studied mouse pre-B-cell differentiation. STATegra includes high-throughput measurements of chromatin structure, gene expression, proteomics and metabolomics, and it is complemented with single-cell data. To our knowledge, the STATegra collection is the most diverse multi-omics dataset describing a dynamic biological system.

Background and Summary

The concept of multi-omics and data-integration has been increasingly used during the last 5 years to describe the multitude of high-throughput molecular technologies that can be applied to the study and analysis of biological systems[1]. Such techniques hold the promise to uncover the different biological processes and layers of regulatory complexity within biological systems. In brief, high-throughput molecular methods can extract information of essentially three basic, yet different components of living cells. Nucleic acids can readily be profiled using massive, parallel sequencing, which in turn provide deep characterization of chromatin properties (ie. Hi-[2]C, ATAC-seq[3], DNase-seq[4], ChIP-seq[5], WGBS[6], RRBS[7]) and the dynamics of gene expression (ie. RNA-seq[8], microRNA-seq[9,10], PAR-CLiP[11], iCLIP-seq[12]). Proteins are measured by proteomics and phosphoproteomics approaches, based on Liquid Chromatography (LC) or Isotope-coded affinity tag labeling (iTRAQ) coupled to Mass Spectrometry (MS). Finally, the metabolome and lipidome, i.e. organic compounds, are captured using mature techniques such as LC/GC-MS or Nuclear Magnetic Resonance (NMR). Increasingly, multi-omics technologies are applied during the same physiological conditions from either the same or different samples to generate a comprehensive set of data spanning multiple molecular levels. The general expectation of multi-omics projects is that the combination of multi-layered data will reveal aspects of the complexity of biological systems that cannot be fully understood using only a particular data-type. Moreover, in

addition to the exciting technical reality of being able to monitor several complementary data-types, the community has come to realize the power of using time in the experimental design. Hence, by collecting data over time, where as a rule the different molecular entities are correlated, it is much more amenable to extract key processes from each data-type as well as uncovering dependencies between different regulatory layers. These technical and conceptual advances are currently being transferred into the vibrant single-cell biology community. Thus, recent advances in single-cell omics technologies have made it feasible to perform multi-omics profiling of individual cells. Consequently, the single-cell community can benefit from the experiences and lessons derived from time-dependent bulk multi-omics analysis. Clearly, a high-resolution single-cell analysis has proven crucial to assess tissue heterogeneity[13-15], cell fate[16,17]. In conclusion, we are most likely entering an era where we can target regulatory networks in single cells[18] using a temporal paradigm coupled to a multi-omics analysis .

While multi-omics projects are frequently depicted as a set of stacked molecular layers that are connected to pass information from the genetic component to the organismal phenotype, the harsh reality is that still many multi-omics project are constrained by budgetary restrictions and sample limitations which evidently reduce the number technologies that can realistically be assessed. In most cases, only a few data types can be included, with a limited number of samples, and analyses is as a rule restricted to focus on 2 or 3 regulatory layers. A few international projects have however successfully collected large datasets and generated comprehensive portfolios of omics measurements. For example, ENCODE[19], TCGA[20], IHE[21], ImmGen[22], had the explicit goal to perform an extensive characterisation of a particular set of cells or tissues. These projects have impacted the scope and type of analysis methods and scientific discoveries that can be achieved so far by the multiomic approach. In some cases combining multi-level data has the ambition to increase the required statistical power to enable the classification of samples or predict disease outcomes. By measuring different types of features. the chance of identifying relevant biomarkers increases, but the analysis does not automatically lend itself to an mechanistic account of the inter-dependencies between these biomarkers as well as their relationship with the outcome, such as a disease. In some cases however, two specific omics layers are measured in order to probe their regulatory relationships. For example, methods that integrate ATAC-seq or RRBS with RNA-seq might shed light on the epigenetic control of gene expression[23], while integrating transcriptomics and metabolomics data may help elucidate metabolic regulation[24,25]. Yet, there have been very few multi-omic studies that evaluate dynamic processes such as cellular growth, development and adaptation. Hence, we still lack formal analysis methods and comprehensive multi-omics datasets that can be leveraged to develop true multi-layered models for dynamic cellular systems. This

state-of-affairs has been the rationale underpinning the formulation of what is referred to as the STATegra project (http://www.stategra.eu/). This is a transnational initiative to develop methods, software and data for dynamic multi-omics analyses. From the STATegra several tools for integrative multiomics data analyses have been published and released[26-32].

Here we share the collection of the different STATegra data-sets, a multi-omics dataset that combines measurements from up to 10 different omics technologies applied to the same biological system. STATegra covers the three types of biomolecules and the different layers that comprise the basic flow of genetic information: chromatin structure (through DNase-seq, RRBS and ChIP-seq), gene expression (RNA-seq and miRNA-seq), proteomics and metabolomics. The collection is complemented with single-cell RNA-seq and ATAC-seq data. STATegra uses a well studied system, namely mouse pre-B-cell differentiation, in a cell line model. This is a highly reproducible *in vitro* system that allows the generation of sufficient material to deploy a comprehensive set of omics measurements. The STATegra multi-omics dataset is unique in the number and diversity of omics technologies available and in the dynamic nature of the system. Our ambition has been to generate this collection of data to serve as gold standard for the development of integrative analysis methods for the multi-layered systems biology.

ChIP-seq data have been used in [33] to identify Ikaros targets. ChIP-seq, DNase-seq, RNA-seq and scRNA-seq datasets were used in [34] to describe the cross-talk between IKAROS Foxo1 and Myc transcription factors in regulating B-cell development.

Methods

Experimental design

Figure 1 illustrates the STATegra dataset. The mouse B3 cell line models the pre-BI (or Hardy fraction C') stage. Upon nuclear translocation of the Ikaros transcription factor these cells progress to the pre-BII (or Hardy fraction D) stage, where B cell progenitors undergo growth arrest and differentiation[33,35]. The B3 cell line was retrovirally transduced with a vector encoding an Ikaros-REt2 fusion protein, which allows control of nuclear levels of Ikaros upon exposure to the drug Tamoxifen[33]. In parallel, cells were transfected with an empty vector to serve as control for the Tamoxifen effect. After drug treatment, cultures were harvested at 0h, 2h, 6h, 12h, 18h and 24hs (Figure 1A) and profiled by several omics technologies: long messenger RNA-seq (mRNA-seq) and small messenger RNA-seq (miRNA-seq) to measure gene expression; reduced representation by bisulfite sequencing (RRBS) to measure DNA methylation; DNase-seq to measure chromatin accessibility as DNaseI Hypersensitive Sites (DHS) and transcription factor footprints, shotgun proteomics

and targeted metabolomics of primary carbon and amino-acid metabolism. Moreover, single cell RNA-seq (scRNA-seq) data for the entire time-series, while bulk ATAC-seq (ATAC-seq) and single-cell ATAC-seq (scATAC-seq) were obtained for 0h and 24h-time points of Ikaros induction only (no control series were run for these datasets). The dataset is complemented by existing ChIP-seq data on the same system equivalent to our 0h and 24h time points[33]. In total, 793 different samples across the different omics datasets define the STATegra data collection (Figure 1B).

Culture conditions

B3 cells containing inducible Ikaros can be expanded before induction of Ikaros to produce sufficient material for all omics experiments. G1 arrest occurs within 16 h following Ikaros induction. Cells containing inducible Ikaros were generated by transducing mouse pre-B cell line B3 with mouse stem cell virus (MSCV) retroviral vectors encoding a fusion protein of haemagglutinin-tagged wild type Ikaros (HA-Ikaros) and the estrogen receptor hormone-binding domain (ERt2), followed by an internal ribosomal entry site (IRES) and GFP. Control cells were generated by transducing mouse pre-B cell line B3 with mouse stem cell virus (MSCV) retroviral vectors encoding the estrogen receptor hormone-binding domain (ERt2) followed by an internal ribosomal entry site (IRES) and GFP. Retroviral infected B3 cells were sorted based on GFP levels. GFP positive cells were expanded in culture for few days (3-4) and then frozen. Frozen vials containing 5 million cells were stored in liquid nitrogen.

For time course experiments, 10 millions of control and Ikaros cells were thawed and expanded for 4 days. Four days later cells were plated for induction of the different time points. Both control and Ikaros cells were split in flasks containing 20 million cells at a density of 0.5 million cells per ml each. For time point inductions, 0.5uM 4-hydroxy-tamoxifen (4-OHT) was added to both, a flask containing Ikaros cells and a flask containing control cells, at one of the specified times: 2h, 6h, 12h, 18h or 24h before collection. Cells for time point 0h (no 4-OHT) induction were obtained separately. All cells were harvested simultaneously. Cells were centrifuged for 5 min at 1200 rpm, washed twice in PBS and counted to aliquot. Aliquots of 10 million cells were done for RNA-seq and metabolomics and proteomics platforms and of 5 million cells for miRNA-seq and Methyl-seq platforms. Cell pellets were snap-frozen in liquid nitrogen and stored at -80. 20-25 million and 50,000 cells were used for DNase-seq and bulk ATAC-seq samples. The full time course experiment was repeated different times (batches) to generate biological replicates (Figure 1B). The same physical cultures were used to obtain cells for mRNA-seq, miRNA-seq, RRBS and proteomics. Other omics tecnologies ran their own cultures to obtain cell material.

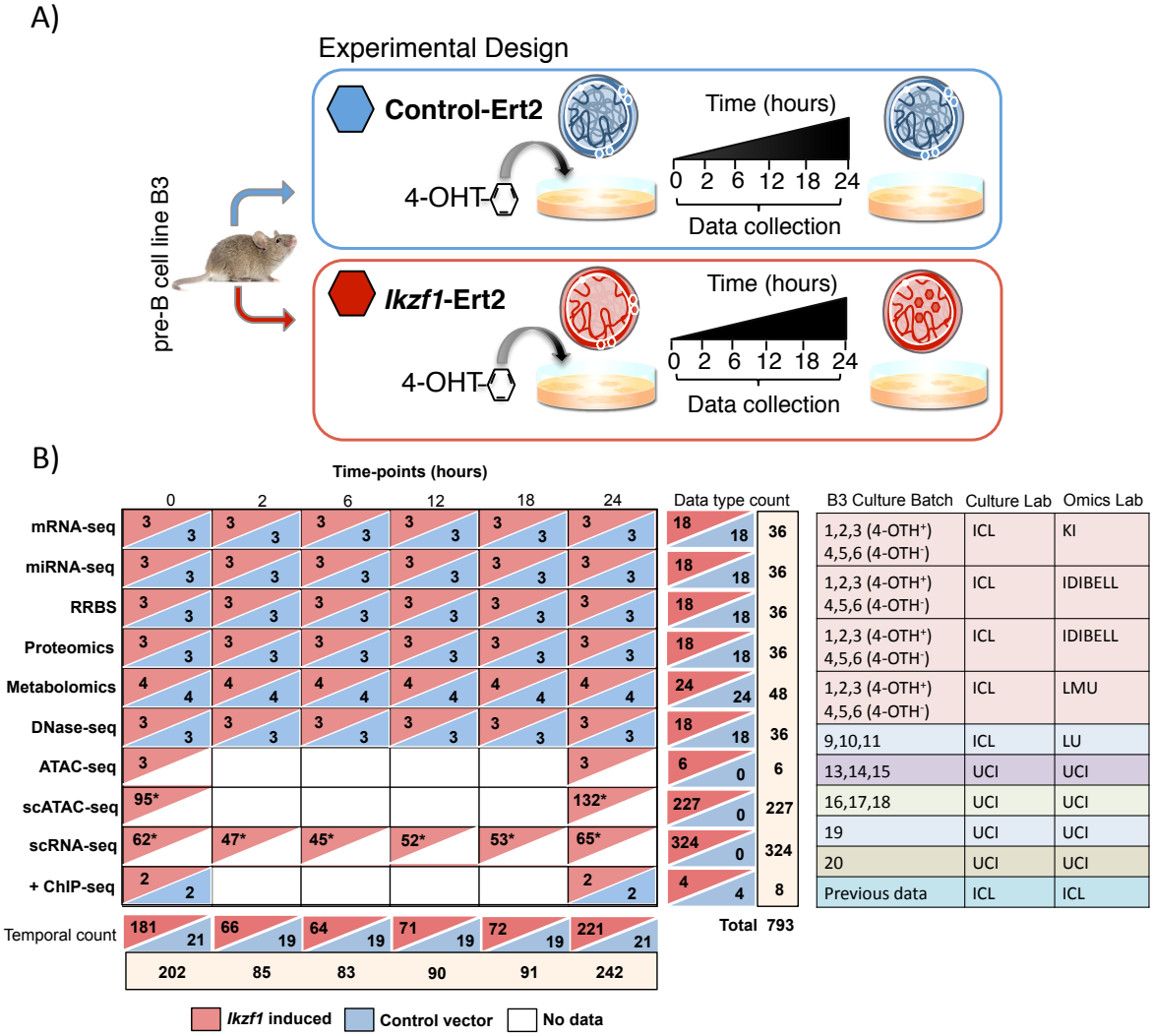


Figure 1. STATegra data generation. A) Inducible Ikaros B3 cell system. Time course experiment collects samples at 6 time points after Tamoxifen induction of Ikaros expression, Control cells carries empty vector. B) Diversity of omics platforms, number of biological replicates, batch distribution and lab assignment for B3 cell culture and omic library perparation. + Previous data from [33]

Aquisition of Multiomics data

RNA-seq

Total RNA was isolated with RNAbee (Ambion), frozen ICL and transported *via courier* (<1 day) to Karolinska Institutet. To account for the impact of the different sources of variability during RNA-seq profiling, we implemented a carefully balanced distribution of samples in relation to time points (6 time points), treatment (Ikaros vs Control), library preparation, bar-code, sequencing run and lanes and biological replicates (3 batches). Briefly, samples were first balanced in six library preparation runs of 6 samples each (Figure 2). Secondly, each RNA-seq library was split into two (total of 72) in order to better account for variability

associated to sequencing. Finally, for sequencing, 75 nucleotides paired-end, the 72 libraries were balanced into 4 flow-cells and in each lane we included 3 libraries. In each lane, we ensured to have different libraries, different batches, different time points and at least both conditions present. Additionally, we balanced the time-points, conditions and batches within each flow-cell. For each flow-cell, a full lane was reserved for quality control. We aimed to obtain 50M reads per library, therefore 100M reads per sample. Libraries were built using the strand-specific RNA-seq dUTP protocol[36]. Sequencing was conducted on an Illumina HiSeq 2500 platform.

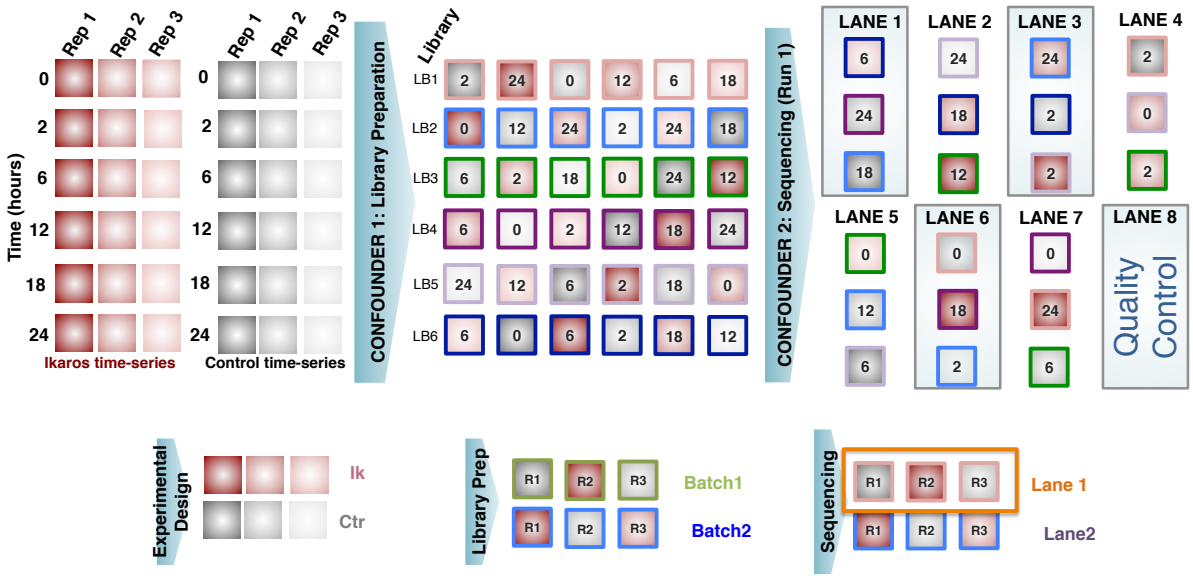


Figure 2. Experimental design for RNA-seq.

miRNA-seq

Small RNA-seq analysis was performed using Trizol-extracted total RNA of 3 biological repicates (4,5,6) for time 0h and total RNA of 3 biological batches (1, 2 and 3) for times 2h, 6h, 12h, 18h and 24h. RNA quality was assessed using Bioanalyzer (Agilent Technologies) evaluating the RNA integrity number (RIN). The library was generated using TruSeq Small RNA Sample Preparation Kit and deep sequencing was performed in Illumina Hiseq 2000 platform. Between 15 and 20 millions of sequencing reads were obtained from each sample.

The library preparation and sequencing of the biological replicates were conducted in two different occasions (technical batches). Figure 3 shows the experimental design according to the batch in which samples were processed. There were two experimental conditions (C=Control, IK=Ikaros) and the 3 biological replicates per condition and time point were numbered as 1, 2 and 3. For some of these biological replicates, also technical replicates were generated in order to estimate the variability between technical batches and to correct any potential batch effect.

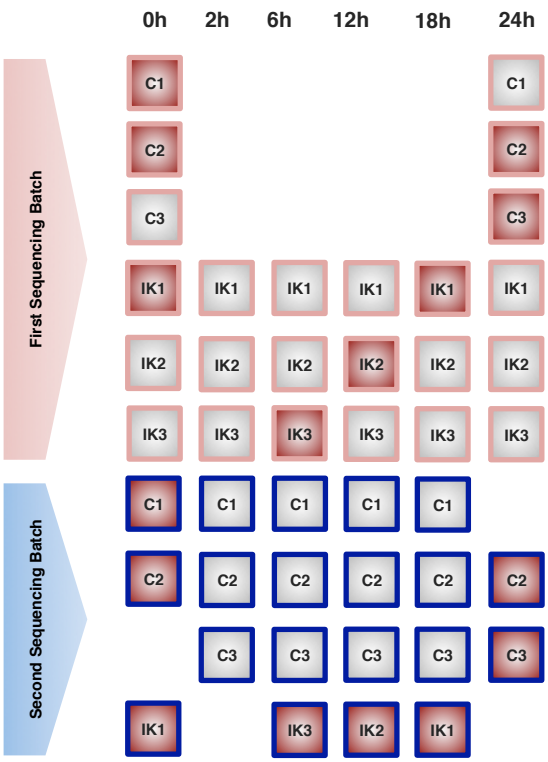


Figure 3. Experimental design for miRNA-seq. Two sequencing batches were run. Samples with red filling were repeated at both batches to allow for estimation of batch effects.

DNase-seq

DNase-seq was performed on ~20-25 million cells with 3 biological replicates for all time-points (0-24 hours) and conditions (Ikaros-inducible and control). Briefly, cells were harvested and washed with cold 1X PBS, prior to nuclei lysis. Lysing conditions were optimized to ensure >90% recovery of intact nuclei. DNaseI concentrations were titrated on Ikaros-inducible and control cells using qPCR against known positive DNaseI hypersensitive promoters (Ap2a1, Ikzf1, Igll1) and negative inaccessible hypersensitive promoters (Myog, Myod) in our biological system, thereby reducing excessive digestion of DNA. Enrichment of DNaseI hypersensitive fragments (0-500bp) was performed using a low-melt gel size selection protocol. Library preparation was performed and sequenced as 43bp paired-end NextSeq 500 Illumina reads. DNaseI libraries were sequenced at a minimum depth of 20 million reads per each biological replicate. To perform DNaseI footprinting analysis, libraries were further sequenced and merged to achieve a minimum of 200 million mapped reads.

RRBS

Genomic DNA was isolated using the high salt method and used for reduced representation bisulfite sequencing (RRBS), a bisulfite-based protocol that enriches CG-rich parts of the genome, thereby reducing the amount of sequencing required while capturing the majority of promoters and other relevant genomic regions. This approach provides both single-nucleotide resolution and quantitative DNA methylation measurements. In brief, genomic

DNA is digested using the methylation-insensitive restriction enzyme MspI in order to generate short fragments that contain CpG dinucleotides at the ends. After end-repair, A-tailing and ligation to methylated Illumina adapters, the CpG-rich DNA fragments (40–220 bp) are size selected, subjected to bisulfite conversion, PCR amplified and then sequenced on an Illumina HiSeq2500 PE 2x100bp[37]. The libraries were prepared for 100-bp paired-end sequencing. Around 30 millions of sequencing reads were obtained from each sample.

Single cell RNA-seq

Single cells were isolated using the Fluidigm C1 System. Single cell C1 runs were completed using the smallest IFC (5-10 um) based on the estimated size of B3 cells. Briefly, cells were collected for each time-point at a concentration of 400 cells/µl in a total of 50 µl. To optimize cell capture rates on the C1, buoyancy estimates were optimized prior to each run. Our C1 single-cell capture efficiency was ~75-90% across 8 C1 runs. Each individual C1 capture site was visually inspected to ensure single-cell capture and cell viability. After visualization, the IFC was loaded with Clontech SMARTer kit lysis, RT, and PCR amplification reagents. After harvesting, cDNA was normalized across all libraries from 0.1-0.3 ng/µl and libraries were constructed using Illumina's Nextera XT library prep kit per Fluidigm's protocol. Constructed libraries were multiplexed and purified using AMPure beads. The final multiplexed single-cell library was analyzed on an Agilent 2100 Bioanalyzer for fragment distribution and quantified using Kapa Biosystem's universal library quantification kit. The library was normalized to 2 nM and sequenced as 75bp paired-end dual indexed reads using Illumina's NextSeq 500 system at a depth of ~1.0-2.0 million reads per library. Each Ikaros time-point was performed once, with the exception of 18 and 24 hour time-points, in which two C1 runs were required in order to achieve approximately ~50 single-cells per each time-point.

Bulk and single cell ATAC-seq

Single-cell ATAC-seq was performed using the Fluidigm C1 system as done previously[38]. Briefly, cells were collected for 0 and 24-hours post treatment with tamoxifen, at a concentration of 500 cells/µl in a total of 30-50 µl. Additionally, 3 biological replicates of ~50,000 cells were collected for each measured time-point to generate bulk ATAC-seq measurements. Bulk ATAC-seq was performed as previously described[3]. ATAC-seq peak calling was performed using bulk ATAC-seq samples. ATAC-seq peaks were then used to estimate single-cell ATAC-seq signal. Our C1 single-cell capture efficiency was ~70-80% for our pre-B system. Each individual C1 capture site was visually inspected to ensure single-cell capture. In brief, amplified transposed DNA was collected from all captured single-cells and dual-indexing library preparation was performed. After PCR amplification of single-cell libraries, all subsequent libraries were pooled and purified using a single MinElute PCR

purification (Qiagen). The pooled library was run on a Bioanalyzer and normalized using Kappa library quantification kit prior to sequencing. A single pooled library was sequenced as 40bp paired-end dual indexed reads using the high-output (75 cycle) kit on the NextSeq 500 from Illumina. Two C1 runs were performed for 0 and 24-hour single-cell ATAC-seq experiments.

Proteomics

A heavy-isotope labeled cell line representing the preB3 cell line at the starting condition was spiked to the sample before trypsin digestion to balance differences in sample amount resulting from sample preparation. After tryptic digestion, proteomic measurements of the 36 biological batches were analyzed by one-dimensional nanoRP-C18 LC-MS/MS in technical triplicates on an LTQ Orbitrap platform coupled to an Ultimate 300 RSLC system (Thermo-Fisher). First, peptide mixtures were desalted on a trapping column (0.3x5mm, Acclaim PepMap C18, 5 µm, Thermo-Fisher) at a flow rate of 25 µl/min of 0.05% TFA. A linear gradient from 3% B to 32% acetonitrile in 0.1% formic acid in 4 h was applied optimal separation of the complete proteome sample. Peptides eluting from the column were directly transferred to the gas phase via a nano-electrospray ionization source (Proxeon) and detected in the mass spectrometer. A data-dependent acquisition cycle consisting of 1 survey scan at a resolution of 60,000 and up to 7 MS/MS scans was employed. Orbitrap MS spectra were internally calibrated on the siloxane signal at 442.1 m/z Charge-state detection was enabled allowing for a precursor selection of charges 2-5 and excluding precursors with undefined, single and higher charge. Precursors with minimal signal intensity of 5000 cps, were isolated within a 1.2 Da window and fragmented by CID (normalized collision energy 35, activation time 30 ms, Q 0.25) and analyzed in the ion trap. Previously analyzed precursors were dynamically excluded from MS/MS selection for 180 seconds.

Metabolomics

Metabolomics measurements were performed on different biological batches than the other omics platforms because the sample preparation part for metabolomics is different than for the rest. In particular, metabolomics requires acute stopping of all metabolic reactions after sampling, while for other types of measurements this is not so critical. Cell extraction protocol for metabolomics consisted of filtration, washing, and quenching steps to remove medium from the cells and stop metabolism. Four biological batches (9, 10, 11 and 12) were acquired. Visual inspection of the cell pellets showed that batch 11 and 12 contained samples which were not completely dry. The metabolomics measurements were obtained with two different analytical platforms, a targeted liquid chromatography mass spectrometry (LC-MS) platform and gas chromatography mass spectrometry (GC-MS) platform. The LC-

MS is a targeted platform measuring amino acids and biogenic amines and the GC-MS focuses on polar metabolites of the primary metabolism such as glycolysis, cyclic acid cycle and amino acid metabolism. LC-MS and GC-MS data had measurements for respectively 36 and 40 metabolites. The measurements were done on exactly the same samples. 80% of the pooled extract was for GC-MS, 10% for LC-MS, 8% for protein weight. Some metabolites were measured at both platforms. In that case the platform that had the lowest relative standard deviation for that metabolite was selected.

The metabolomics measurement pipeline includes two types of control: the quality control (QC) sample and the internal standard solution. The QC sample is typically a mixture of study samples that is inserted after each 6 study samples in the measurement series and is used to correct for experimental drift of the analytical instrument. Because of limited availability of sample material the QC sample used here was not a mixture of study samples but material of control B3 cells not activated with tamoxifen. The internal standard solution for the GC-MS and LC-MS consists of 13C labelled yeast extract which is added to each study sample at the beginning of the sample preparation process to correct for experimental errors made during the sample processing. For LC-MS an additional internal standard solution is added consisting of 13C labelled amines for most of the amines measured with the platform. For LC-MS the labelled versions of the metabolites were used as internal standard while for GC-MS the best internal standard was chosen based on the smallest residual standard deviation of the QC samples. During the process of measurement, the time points for each batch are randomized, but each Ikaros sample and its control is maintained together.

Omics pre-processing

Data pre-processing is next described in detail for each omics type. Figure 4 shows an comparative overview of the different preprocessing pipelines.

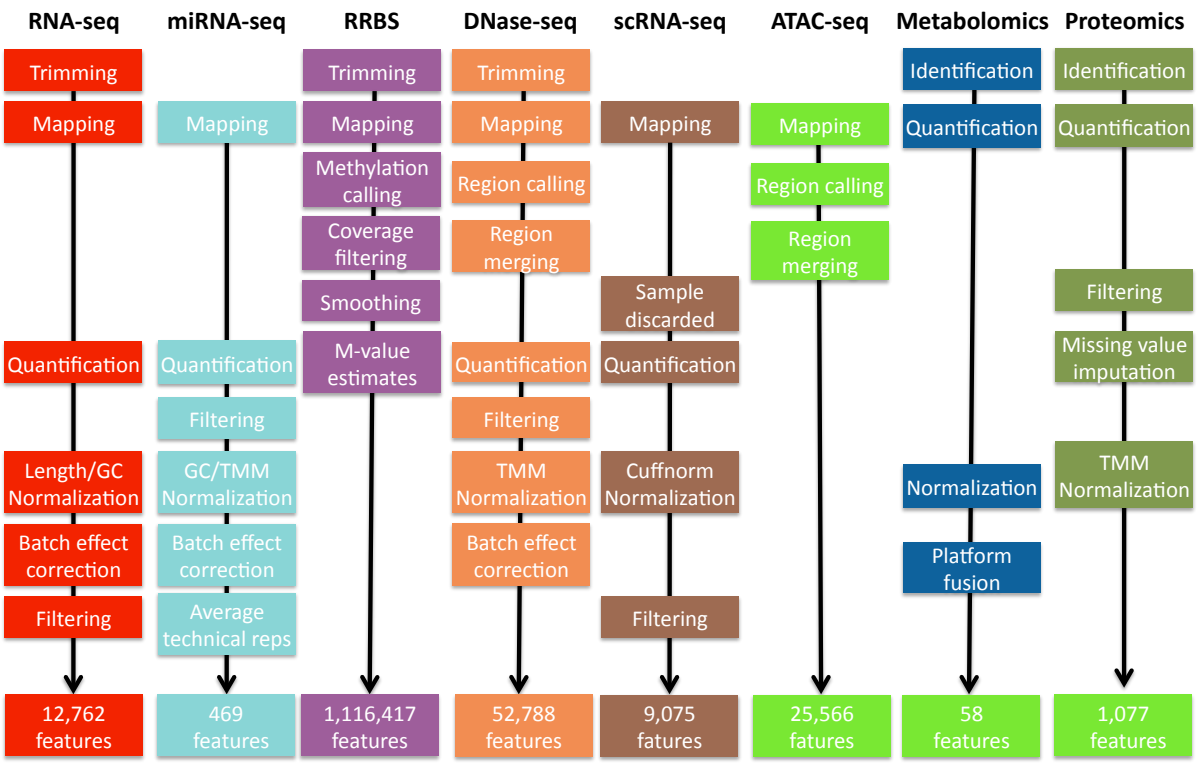


Figure 4. Preprocessing pipelines for 7 omics technologies. See methods for details.

RNA-seq

Tophat2[39] was used to map fragments to the mm10 reference genome; the very-sensitive mode only allowing a unique best mapping per fragment was used. Piccard (https://broadinstitute.github.io/picard/) and Fastqc[40] were used to perform a quality control considering elements such as duplication levels, GC content and k-mer overrepresentation. We observed that the duplication level was high (over 90%) in most samples as expected for high sequencing-depth in RNA-seq; additionally, few samples were having a GC content over-representation (results not shown). Trimming was applied to remove Illumina primers and low-quality nucleotides. HTSEQ[41] *intersection-option* was used to assign fragments to genes. Data were normalized using *cqn*[42], which corrects for GC content and gene-length. A non-parametric version of Combat methodology[43] was used after cqn to correct for library-preparation effects.

miRNA-seq

The quality of the sequencing reads was checked with the Fastqc tool with good results (results not shown). Alignment of raw data was performed using Novoalign (http://novocraft.com/) on mouse miRNA sequences from mirBase. Quantification was performed using multiBamCov[44] and counts were found for 1,086 out of 1,908 miRNAs present in the database. Low count miRNAS were further filtered out with the CPM (counts per million) method in NOISeq R package[45] by setting a threshold of 1 for CPM. The final

dataset contained 469 miRNAs. GC content bias, was eliminated with the cqn R package[42], data was normlized by TMM[46] and batch effects corrected by ComBat[43]. After batch correction, technical replicates were averaged for further analyses.

DNase-seq

DNase-seq reads were trimmed to 36bp and paired-end mapped to the mm10 reference genome using Bowtie2[47] with options: -v 2 -k 1 -m 1 --best –strata. DNase-seq peaks were called for each replicate using the HOMER findPeaks function. We employed a specific peak-calling strategy to capture several features of our DNaseI hypersensitive sites. Our strategy was to include both 'narrow' and 'broad' DHS peaks in our analysis. This captured a comprehensive set of sites with a wide DHS dynamic range. Initially, we used HOMER to determine narrow DHS peaks using a default size parameter (120-150bp) with a minimum peak distance of 50bp between DHS and an FDR of 1%. We then included a second round of peak calling, restricting to a peak size of 500bp with a minimum peak distance of 50bp between DHS peaks and an FDR of 1%. We then merged the two peak sets for each replicate. We required a minimum 1bp overlap of peaks across all three biological replicates for each time-point respectively and generated a consensus DHS peak list across all time-points.

The consensus DHS (53,624) were filtered for chrM peaks, partial chromosomes, and mouse ENCODE blacklist regions. Counts representing the chromatin accessibility were estimated for each consensus DHS using the Bedtools coverageBed function. Additionally, no DHS were considered with less than 10 reads (~1 RPM) in all time-points, resulting in a final dataset with 52,788 consensus DHSs. Data were normalized by a combination of RPKM and TMM and batch effects were removed by the ARSyN method[48].

RRBS

Initial quality assessment was based on data passing the Illumina Chastity filter. The second quality assessment was based on the reads using the Fastqc quality control tool version 0.10.0. Reads were adaptor- and quality-trimmed using Trim-Galore Software v0.3.4 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) in RRBS paired-end mode, in order to decrease methylation call errors arising from poor quality data. Mapping to the reference genome (GRCm38, mm10) was performed using Bismark v0.10.1[49] and Bowtie2[47]. The quality of the mapping was inspected using HTSEQ-qa[41]. SAM files were used as input in Bismark to obtain methylation calls. Paired-end mode with no overlap mode was specified. The first four bases from each read were avoided to eliminate M-bias, i.e. deviation from the horizontal line in the mean CpG methylation level for each read position. BedGraph and *.cov files were further considered and analyzed with the BiSeq package[50].

Coverage was inspected before proceeding to smooth the methylation levels (between 0 and 1) per CpG site. Briefly, we firstly defined “frequently covered CpG sites” as those sites that are covered in at least 2/3 of the samples. The frequently covered CpG sites were considered only to define the cluster boundaries and we defined CpG clusters using a maximum distance of 100 bp and at least 20 CpGs. This selection resulted in 1,116,417 CpG sites within CpG clusters, with no threshold on coverage. The extra coverage of unusually high covered sites (95% quantile of the coverage) was eliminated to remove potential biases during the smoothing step introduced by CpGs with exceptional high coverage. Then, the methylation levels were smoothed with a bandwidth of 80 bp as described[50]. Clustering analysis was performed with the methylation estimates of the 20% most variable positions (based on CV), with multidimensional scaling or hierarchical clustering. M-values were obtained after thresholding methylation levels in the interval [0.01, 0.99] to avoid infinite values, as $M = \log2(b/(1-b))$, where b is the constrained methylation level. The final dataset contained a total of 1,116,417 Methylation features.

Single cell RNA-seq

A total of 560 single-cell RNA-seq libraries were mapped with Tophat[51] to the mouse Ensembl gene annotations and mm10 reference genome. Single-cell libraries with a mapping rate less than 50% and less than 450,000 mapped reads were excluded from any downstream analysis, resulting in 324 single-cells for all subsequent analysis. Cufflinks[52] version 2.2.1 was used to quantify expression from single-cell libraries using Cuffquant. Gene expression measurements for each single-cell library were merged and normalized into a single data matrix using Cuffnorm. Genes with zero counts in more than 80% of the samples were removed resulting in a data matrix with 9,075 genes.

ATAC-seq

Single-cell libraries were mapped with Bowtie[47] to the mm10 reference genome using the following parameters (bowtie -S -p 2 --trim3 10 -X 2000). Duplicate fragments were removed using Picard (http://picard.sourceforge.net). We considered single-cell libraries that recovered > 5k fragments after mapping and duplication removal. Bulk ATAC-seq replicates were mapped to the mm10 reference genome using the following parameters (bowtie2 -S -p 10 --trim3 10 -X 2000). Peak calling was performed on bulk replicates using HOMER with the following parameters (findPeaks <tags> -o <output> -localSize 50000 - size 150 -minDist 50 –fragLength 0). The intersection of peaks in three biological replicates was performed. A consolidated list of 25,466 peaks was generated from the union of peaks from 0 and 24 hour time-points.

Proteomics

Data were searched against a protein sequence database containing all confirmed mouse protein sequences from the Uniprot database (swissprot), common contaminants and reversed using the Andromeda algorithm within the MaxQuant software suite version 1.5.0.0. Mass deviation settings for peptide detection were 20 ppm for the first search and 7 ppm for the main search. IT MS/MS data were searched with a mass accuracy of 0.6 Da. N-terminal acetylation and methionine oxidation were set as variable modifications, carbamidomethylation of cysteine as fixed modification. Unidentified signals, present at a similar retention environment, were matched if at least one run had a positive identification of the peptide sequence by enabling the match between run option within an alignment time window of 30 min. Obtained search results were filtered for reversed database hits on the peptide spectrum match level (1%) and the protein level (2%) and all protein groups with at least 1 razor or unique peptide were initially accepted. For quantitation of proteins over the different timepoints, light protein intensities were extracted from the data file for each protein.

Distant measuring intervals and long gradient times lead to a substantial variation between peak localization and areas of individual LC-MS/MS runs resulting in a huge number of missing values. This issue was addressed by:

a) Alternative LC MS/MS alignment routines. Upon the observation that missing values were not randomly distributed, but associated to particular samples, we believed that a misalignment of chromatograms played a role. We improved the alignment between samples to rescue some of the missing values.
b) RNA-seq data was used as database source protein identification. The rationale is that the mRNAs of expressed proteins should be found within the RNA-seq detected genes. Reducing the size of the protein database to proteins detected by RNA-Seq will reduce the number of false hits and lead therefore to more specific data on the pre-B cell proteome. Since proteomic data exhibit a lower coverage of the proteins abundant in a cell, a substantial data loss is not expected.
c) Missing value imputation was applied on log2-transformed data corresponding to 2,527 proteins. Those proteins with missing values in all the 3 replicates per condition in at least 11 of the 12 conditions were discarded, resulting in 2,396 proteins used for the imputation. Protein with all missing values in the Control condition were imputed from a gaussian distribution with mean 0 and with standard deviation equal to the median of all within-group standard deviations for all the proteins in the original data. The same procedure was followed for Ikaros condition. Next, the algorithm “norm.predict” from the MICE R CRAN package (version 2.3) was applied with default values to impute the rest of missing values in proteins with at

least one non-missing value per experimental conditions. Proteins with three missing values per conditions were replaced by NAs for all values in that condition.

The resulting imputed data with no missing values (1,077 proteins) were normalized by TMM. Additionally, a subset of *bona fide* proteomics values was defined as those proteins with measurements in at least 34 of the 36 samples, resulting in 216 proteins.

Metabolomics

For the GC platform, a 13C labelled yeast extract was added as internal standard. A Quality Control (QC) sample was measured every 6 samples. Of each of the compounds measured on the GC platform, the labelled compound that lead to the smallest standard error in the QC samples for that compound, was selected as internal standard. Because the amount of sample was almost completely used for the two analytical platforms no replicate analyses were possible. This meant that the internal standard selection could not be validated using replicate samples as is common practice. For the targeted LCMS method, the optimal internal standards for each metabolite were chosen during optimization and validation of the method. The limited within batch drift effects were corrected using the batch correction approach developed by van der Kloet *et al*[53].

Four biological batches (batches 9 through 12) were provided to the metabolomics platforms, which were (physically) different from the batches used for mRNA-seq, miRNA-seq and proteomics. Visual inspection showed that samples of batch 11 and 12 were not completely dry. Analysis of some key metabolites showed batch 12 levels to be outlying the general trend in batches 9, 10 and 11. Therefore, it was decided to exclude batch 12 from further analysis.

Both analytical platforms show some overlap in the metabolites that were measured. On GCMS 22 metabolites were uniquely quantified and 18 metabolites were quantified uniquely on LCMS, while 18 metabolites were quantified both on GCMS and LCMS, making a total of 58 metabolites. Although the intensity levels of the GCMS and LCMS were rather different a high correlation between the two platforms for most overlapping metabolites was observed. Metabolite levels were log scaled and levels were mean-centered over the three batches 9, 10 and 11. For the metabolites that were measured both on GCMS and LCMS weighted averages were calculated where the weight was defined as the inverse of the relative standard deviation (RSD) of that metabolite on its specific platform obtained from repeated analyses of pooled samples.

Data Records

Raw data

STATegra multiomics data have been deposited in different public repositories dedicated to the different data types. Table 1 shows a list of the current hosting of raw data files. Moreover, pre-processed data arranged as a data-matrix per omics data-type have also made available at Lifebit (https://lifebit.ai/) site.

Table 1. Public repositories hosting STATegra multi-omics data.

Data set	Database and accession
mRNA-seq	GEO, GSE75417
miRNA-seq	GEO, GSE75394
RRBS	GEO, GSE75393
DNAse-seq	GEO, GSE75390
ATAC-seq	GEO, GSE89362
scRNA-seq	GEO, GSE89285
scATAC-seq	GEO, GSE89362
ChIP-seq	GEO, GSE38200
Proteomics	ProteomeXchange, PXD003263
Metabolomics	MetaboLights, MTBLS283

STATegra Knowledge Base

In order to evaluate how to best integrate and semantically map specific prior knowledge and relevant information derived from multiple sources together with heterogeneous experimental data, a STATegra Knowledge network for B-cell differentiation (KB) was developed (applying the BioXM™ knowledge management environment[54] (https://ssl.biomax.de/stategrakb/). Prior knowledge includes among others relevant molecular elements (genes, proteins, metabolites, etc.), functional information (GO, OMIM, etc.), functional interactions (e.g. protein-protein interaction, transcriptional regulation (e.g. mouse TF-regulatory network), miRNA network, etc.) and information about gene homologs (mouse, rat, human). Also, genome features with coordinates for peak to gene associations of NGS data (e.g. mouse genome assembly mm10), metabolic and signal transduction pathways, cell types related to B-cell differentiation as well as ontologies such as the mouse anatomy ontology (MGI) were incorporated. This integrated and dynamically organized knowledge serves as information rich, structured background network. Semantic mapping of experimental data to this background network of prior knowledge enables complex

integration and analysis approaches[55]. From this project specific solution for integrated analysis we abstracted specific open tools for data access presented in this manuscript.

The STATegra Knowledge Base visualizes multiple omics data types and summarizes information from different layers on top of a network graph. An example for a dynamic visualization of several different experimental data sources on top of a background network based on public knowledge is shown in Figure 5. The overlay of own experimental data on top of such networks would help interpretation of results as well as validate database predictions.

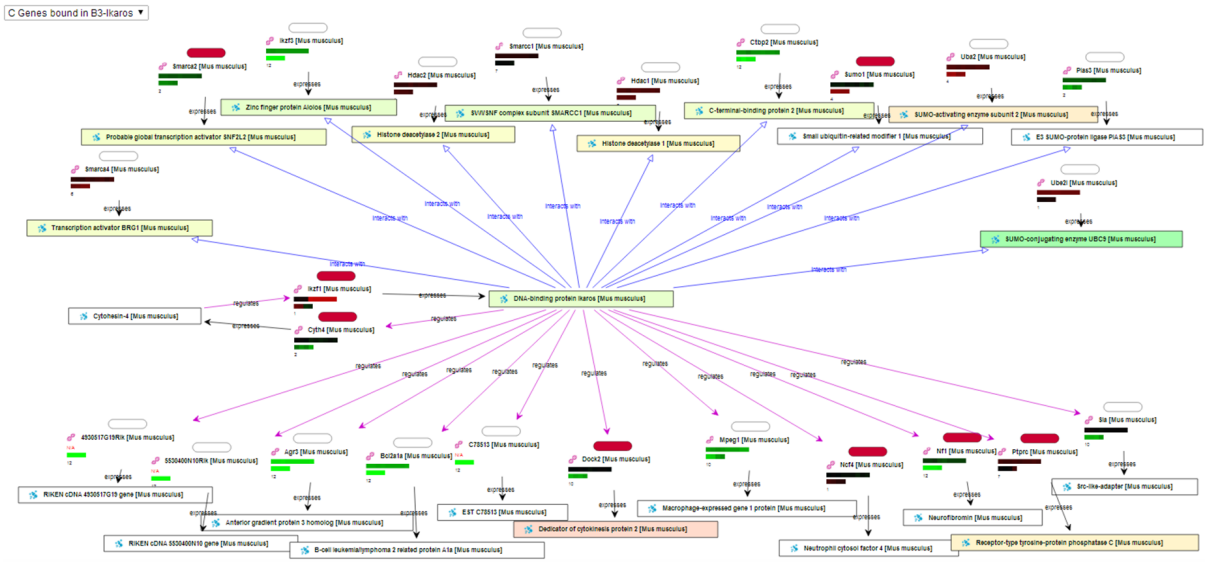
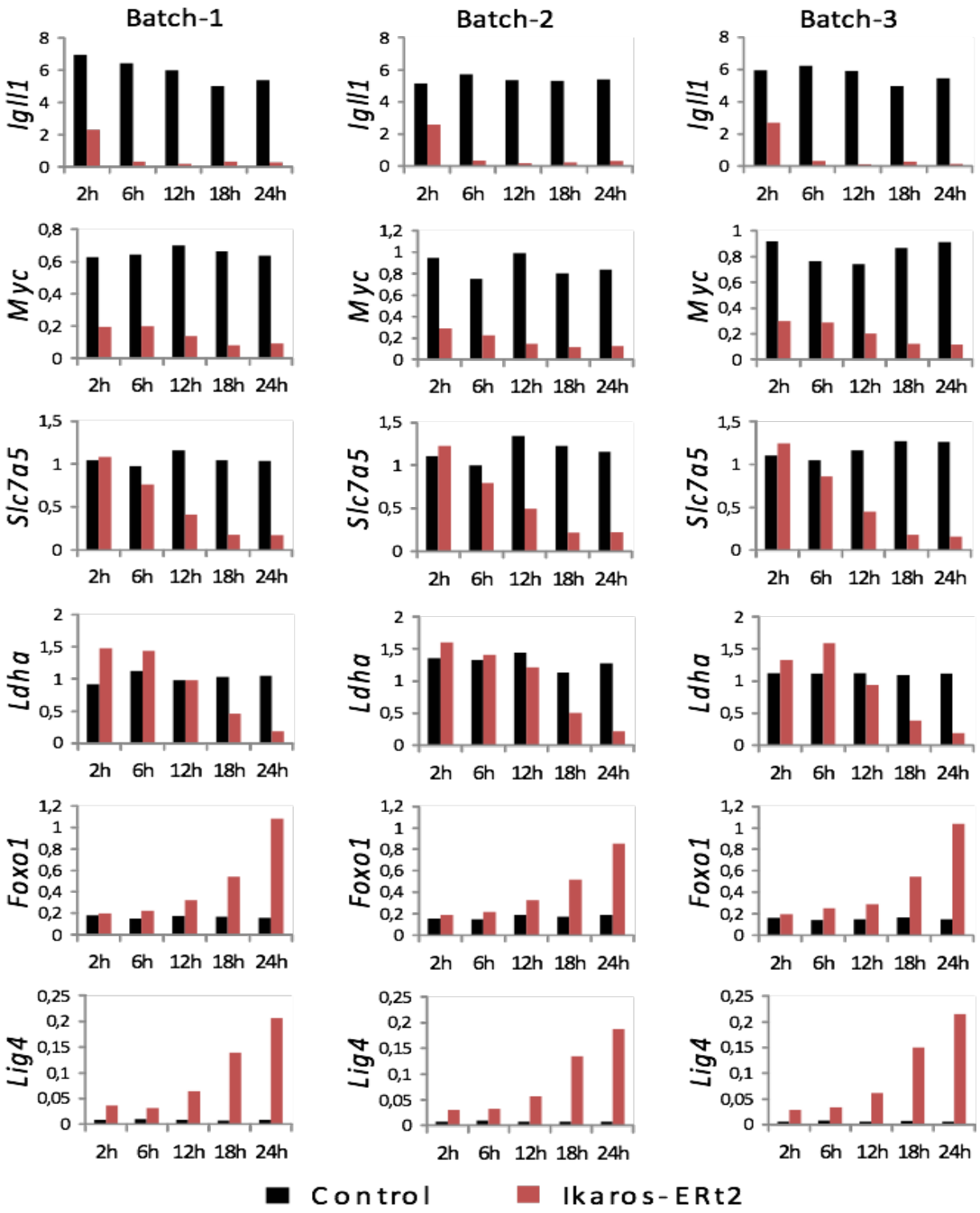


Figure 5. Example of a dynamic visualization of several different experimental data sources on top of a background network based on public knowledge. The Ikaros-Network is based on direct protein-protein interactions (BioGRID, MINT, IntAct; pink arrows) and predicted TF-target relations (ITFP[56]; blue arrows). Visualization includes a series of gene expression experiments (upper and lower heatmap rows correspond to GSM934758-61 and GSM934762-70, respectively), gene clusterings (small numbers below heat maps indicate expression based clusters),ChIP-seq results (red ellipses on top of the genes indicate that a peak overlapping the corresponding gene region was identified in both B3Ikaros and B3HA-Ikaros Chip-seq experiments (GSE38169). In addition, protein quantification values are overlaid with the protein nodes (rectangles, in this case only example data from control cells (mouse pre B3 cell line) are shown). Colors and nodes are freely configurable.

Technical validation

Validation of time course replicability

As a quality control of batch replicability, real time RT-PCR was used to check the impact of Ikaros in gene expression upon induction and reproducibility across time course experiments. RNA from all samples was extracted using RNAbee (AMS Biotechnology (Europe) Ltd) and treated with Turbo DNAse (Life technologies). Bioanalyser technology was used to check the RNA integrity and samples were quantified using a Nanodrop. Changes in the expression of few previously identified Ikaros-responsive genes were analysed (Figure



6). As expected, early down-regulation of *Igll1* and *Myc*, late down-regulation of *Slc7a5*, *Hk2* and *Ldha*, and up-regulation of *Foxo1* and *Lig4* were consistently observed in the three independently collected time course replicates. Either frozen pellets or RNA samples from the time course experiments and 0h time point collections were sent to the different experimental labs to perform the library preparation for the sequencing.

Figure 6. **Biomarkers of B3 cell differentiation across three experimental batches.**

Validation of dataset co-variation structure.

To understand wether the different omics measurements captured the dynamics of B-cell differentiation and/or had a similar co-variation structure, we ran Principal Component Analysis for all datasets (Figure 7). In general, the different multiomics datasets show PCA plots that recapitulate the time progression of our inducible system. A well spread temporal progression on the first PC was observed for mRNA-seq, miRNA-seq and scRNA-seq data, being RNA-seq the dataset with the most consistent progression signal. Metabolomics and Proteomics showed a two stage pattern, with samples from 0h-12h hours clustering at negative values, and samples at 18h-24h clustering at positive values of the first and second PC, respectively. DNase-seq showed a noisier, but distinguible distribution of the temporal signal at PC2, while RRBS is the only dataset with an unclear temporal pattern. For scATAC-seq, only two time points were measured and cells nicely separated on the second PC. This analysis reveals that multi-omics datasets consistently described the progression of B-cell differentiation but also that the different omics techologies present different noise levels. Interestingly, dynamic patterns are slightly different for nucleic acids and proteins and metabolites, possibly indicating a later reponse of these with respect to the transcriptional change.

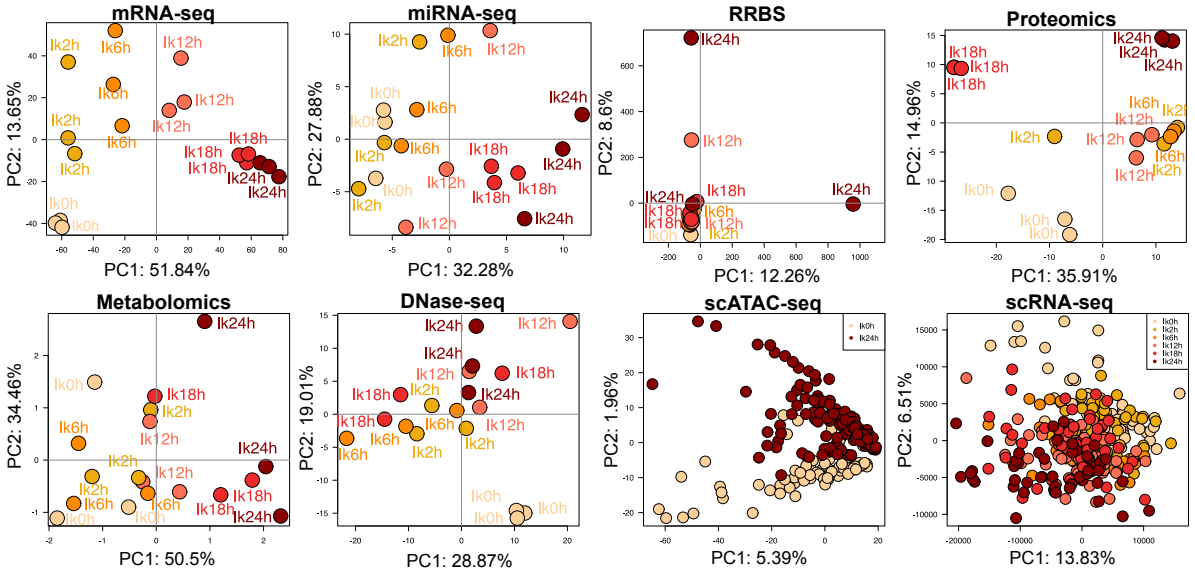


Figure 7. PCA analysis of STATegra multiomics data. Only the Ikaros series is shown. Data was preprocessd as described in Methods. Time progression is represeting by an increasingly darker red color.

Acknowledgements

This work has been funded by the European Union Seventh Framework Programme [FP7/2007–2013] under the grant agreement 306000-STATegra.

The author declare no competing interests

Author contributions

DGC: Conceived the study, contributed to RNA-seq data generation, performed preprocessing and validation of multiple omics datasets

ST: Performed preprocessing of multiple omics datasets, created initial draft of the paper

IFV: Generated B3 cell line cultures for multiomics profiling

RNR: Performed DNase-seq, ATAC-seq, scATA-seq and scRNA-seq experiments. Contributed to data pre-preocessing and drafting of the manuscript.

CC: Performed RRBS and microRNA-seq experiments

AS: Performed Proteomics experiments and contributed to data preprocessing

TR: Performed Metabolomic experiments and contributed to data preprocessing

VvSP: Developed STATegra KnowledgeBase portal

FM: Performed preprocessing of RRBS data

JRU: Contributed to the preprocessing of miRNA-seq and RRBS data

AGG: Contributed to the preprocessing of miRNA-seq and RRBS data

TC: Contributed to generating B3 cell line cultures for multiomics profiling

LC: Contributed to generating B3 cell line cultures for multiomics profiling

ZL: Contributed to generating B3 cell line cultures for multiomics profiling

GD: Contributed to generating B3 cell line cultures for multiomics profiling

LBN: Contributed to the analysis of multiomics dataset

VL: Contributed to data analysis

IS: Contributed to supervision of analysis

ML: Contributed to supervision of analysis

DM: Supervised STATegra KnowledgeBase development

JAW: Supervised data preprocessing

TH: Supervised Metabolomics data generation

AI: Supervised Proteomics data generation

EB: Supervised miRNA-seq and RRBS data generation

AM: Supervised DNase-seq and single cell data generation

MM: Supervised cell culture experiments and validation

JT: Conceived the study and supervised analyses

AC: Conceived and coordinated the study supervised analyses and drafted the manuscript.

References

1 Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC Systems Biology* **8**, I1, doi:10.1186/1752-0509-8-S2-I1 (2014).

2 Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods (San Diego, Calif.)* **58**, 268-276, doi:10.1016/j.ymeth.2012.05.001 (2012).

3 Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology* **109**, 21.29.21-21.29.29, doi:10.1002/0471142727.mb2129s109 (2015).

4 Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**, pdb prot5384, doi:10.1101/pdb.prot5384 (2010).

5 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).

6 Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature genetics* **47**, 746-756, doi:10.1038/ng.3291 (2015).

7 Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* **33**, 5868-5877, doi:10.1093/nar/gki901 (2005).

8 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).

9 Juzenas, S. *et al.* A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic acids research* **45**, 9290-9301, doi:10.1093/nar/gkx706 (2017).

10 Jima, D. D. *et al.* Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* **116**, e118-e127, doi:10.1182/blood-2010-05-285403 (2010).

11 Hafner, M. *et al.* PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins. e2034, doi:doi:10.3791/2034 (2010).

12 Konig, J. *et al.* iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *Journal of visualized experiments : JoVE*, 2638, doi:10.3791/2638 (2011).

13 Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661 (2017).

14 Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776, doi:10.1126/science.1247651 (2014).

15 Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).

16 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386, doi:10.1038/nbt.2859 (2014).
17 Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology* **14**, e8046-e8046, doi:10.15252/msb.20178046 (2018).
18 Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817, doi:10.1016/j.cell.2016.11.038.
19 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, doi:10.1371/journal.pbio.1001046 (2011).
20 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).
21 Stunnenberg, H. G. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1897, doi:10.1016/j.cell.2016.12.002 (2016).
22 the ImmGen, C. & Benoist, C. Open-source ImmGen: mononuclear phagocytes. *Nature Immunology* **17**, 741, doi:10.1038/ni.3478 (2016).
23 Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e1324, doi:https://doi.org/10.1016/j.cell.2016.10.026 (2016).
24 Dumas, M.-E. *et al.* Topological analysis of metabolic networks integrating co-segregating transcriptomes and metabolomes in type 2 diabetic rat congenic series. *Genome medicine* **8**, 101-101, doi:10.1186/s13073-016-0352-6 (2016).
25 Mastrokolias, A. *et al.* Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics* **12**, 137, doi:10.1007/s11306-016-1084-8 (2016).
26 Furió-Tarí, P., Conesa, A. & Tarazona, S. RGmatch: matching genomic regions to proximal genes in omics data integration. *BMC Bioinformatics* **17**, 1-10, doi:10.1186/s12859-016-1293-1 (2016).
27 Hernández-de-Diego, R. *et al.* PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Research* **46**, W503-W509, doi:10.1093/nar/gky466 (2018).
28 Hernandez-de-Diego, R. *et al.* STATegra EMS: an Experiment Management System for complex next-generation omics experiments. *BMC Syst Biol* **8 Suppl 2**, S9, doi:10.1186/1752-0509-8-S2-S9 (2014).
29 Nueda, M. J., Tarazona, S. & Conesa, A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* **30**, doi:10.1093/bioinformatics/btu333 (2014).
30 Reshetova, P., Smilde, A. K., van Kampen, A. H. C. & Westerhuis, J. A. Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. *BMC systems biology* **8 Suppl 2**, S2-S2, doi:10.1186/1752-0509-8-S2-S2 (2014).
31 Karathanasis, N., Tsamardinos, I. & Lagani, V. omicsNPC: Applying the Non-Parametric Combination Methodology to the Integrative Analysis of Heterogeneous Omics Data. *PloS one* **11**, e0165545-e0165545, doi:10.1371/journal.pone.0165545 (2016).
32 Martínez-Mira, C., Conesa, A. & Tarazona, S. MOSim: Multi-Omics Simulation in R. *bioRxiv*, 421834, doi:10.1101/421834 (2018).
33 Ferreiros-Vidal, I. *et al.* Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. *Blood* **121**, 1769-1782, doi:10.1182/blood-2012-08-450114 (2013).
34 Isabel Ferreirós Vidal, T. C., Vincenzo Lagani, Ricardo Ramirez, Elizabeth, Ing-Simmons, A. G., Lee Cooper, Ziwei Liang, Georgios, Papoutsoglou, G. D., Sonia

Tarazona, Sunjay J. Fernandes, Peri, Noori12, G. S., Amanda G. Fisher, Ioannis Tsamardinos, Ali Mortazavi, & Boris Lenhard, A. C., Jesper Tegner, Matthias Merkenschlager, David Gomez-Cabrero. Systematic identification of transcription factors in B cell progenitor differentiation. *PloS Biol* **In press** (2019).

35 Hardy, R. R., Carmack, C. E., Shinton, S. A., Kemp, J. D. & Hayakawa, K. Pillars Article: Resolution and Characterization of Pro-B and Pre-Pro-B Cell Stages in Normal Mouse Bone Marrow. J Exp Med. 1991. 173:1213–1225. *The Journal of Immunology* **189**, 3271 (2012).

36 Borodina, T., Adjaye, J. & Sultan, M. in *Methods in Enzymology* Vol. 500 (eds Daniel Jameson, Malkhey Verma, & Hans V. Westerhoff) 79-98 (Academic Press, 2011).

37 Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols* **6**, 468, doi:10.1038/nprot.2010.190 (2011).

38 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, doi:10.1038/nature14590 (2015).

39 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, doi:10.1186/gb-2013-14-4-r36 (2013).

40 Andrews S. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 29 September 2014.

41 Anders, S., Pyl, P. T. & Huber, W. HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, doi:10.1093/bioinformatics/btu638 (2015).

42 Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics (Oxford, England)* **13**, 204-216, doi:10.1093/biostatistics/kxr054 (2012).

43 Stein, C. K. *et al.* Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC bioinformatics* **16**, 63-63, doi:10.1186/s12859-015-0478-3 (2015).

44 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **26**, doi:10.1093/bioinformatics/btq033 (2010).

45 Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* **43**, e140, doi:10.1093/nar/gkv711 (2015).

46 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, doi:10.1186/gb-2010-11-3-r25 (2010).

47 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, doi:10.1038/nmeth.1923 (2012).

48 Nueda, M. J., Ferrer, A. & Conesa, A. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* **13**, 553-566, doi:10.1093/biostatistics/kxr042 (2012).

49 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).

50 Klein, H.-U., Hebestreit, K. & Dugas, M. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29**, 1647-1653, doi:10.1093/bioinformatics/btt263 (2013).

51 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105-1111 (2009).

52 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

53 van der Kloet, F. M., Bobeldijk, I., Verheij, E. R. & Jellema, R. H. Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. *Journal of Proteome Research* **8**, 5132-5141, doi:10.1021/pr900499r (2009).
54 Losko S , H. K. in *Protein Networks and Pathway Analysis. Methods in Molecular Biology* Vol. 563 (ed Bryant J Nikolsky Y) (Humana Press, 2009).
55 Maier, D. *et al.* Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC systems biology* **5**, 38-38, doi:10.1186/1752-0509-5-38 (2011).
56 Zheng, G. *et al.* ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics* **24**, 2416-2417, doi:10.1093/bioinformatics/btn439 (2008).