PROJECT FINAL REPORT

Grant Agreement number: 305626

Project acronym: RADIANT

Project title: Rapid development and distribution of statistical tools for high-throughput sequencing data

Funding Scheme: Collaborative Project (FP7-HEALTH-2012-2.1.1-3: Statistical methods for collection and analysis of -omics data)

Period covered: from 1.12.2013 to 30.11.2015

Name of the scientific representative of the project's co-ordinator[1], Title and Organisation:

Prof. Magnus Rattray. The University of Manchester

Tel: 0161 275 5094

Fax:

E-mail: Magnus. Rattray@manchester.ac.uk

Project website address: http://radiant-project.eu

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement.

4.1 Final publishable summary report

4.1.1 Executive summary

The objectives of the RADIANT project were to develop improved computational tools for the analysis of high-throughput sequencing (HTS) data, to rapidly disseminate these tools to the wider scientific community and to support practitioners in the application of these and other advanced HTS data analysis tools in their scientific applications, lowering the bar for the widespread adoption of cutting-edge statistical tools. The project has been successful in achieving these objectives.

The RADIANT project supported the development of several superb data analysis tools which have been widely adopted by practitioners. The new DESeq2 tool for differential expression analysis was developed entirely through RADIANT funding and provides substantial improvements over the earlier DESeq tool; it is one of the most widely-used tools in genomics (Love et al., 2014 cited 228 times in Scopus, 34500 software downloads in the last year). Similarly, the well-established edgeR and HTSeq tools are extremely popular tools which have been enhanced through RADIANT research developments. Other new developments supported by RADIANT include improved tools for transcript-level expression estimation (BitSeqVB, Hensman et al. 2015, most read paper in Bioinformatics, Dec. 2015); new methods for transcript structure inference (FlipFlop); new tools for epigenomic data analysis (Repitools, CompEpiTools) and new tools for HiC data analysis (Pastis, Centurion, HiC-Pro, FourCSeq). A visualisation tool, the RADIANT Genome Browser, was developed to visualise the data and results from these HiC and epigenomics analysis tools. RADIANT partners have also developed methods for the model-based analysis of time course data from studies of transcription dynamics (DETime; INSPEcT) to better understand transcriptional regulation.

Most of the above tools have been included in R/Bioconductor providing a uniform framework for HTS data analysis, method development, distribution and documentation. However, even within Bioconductor, data containers may differ across packages making inter-operability non-trivial. The popularity of Bioconductor also means that there are a huge number of packages for users to navigate, making is difficult for the inexperienced user to solve a specific data analysis problem. Therefore, a major focus of the RADIANT project has been on improving the quality of documentation (e.g. through the BiocStyle package), inter-operability of packages (through the development of converter packages) and providing task-oriented rather than tool-oriented documentation for users (e.g. the “Bioconductor channel” in the F1000research journal).

HTS technologies are evolving rapidly and the RADIANT project has taken measures to stay up-to-date and develop cutting-edge statistical tools for important new datatypes. We have supported scientific meetings investigating methods for new technologies such as long-read and single-cell sequencing, and we have applied and developed methods in these areas. For example, the IONiseR tool provides a much-needed resource for quality control of data from the increasingly popular MinION long-read technology.

An important aspect of method development is appropriate benchmarking. Throughout the project we have dedicated significant effort to this topic through project meeting sessions, RADIANT-sponsored workshops and collaborative activities. In this context, Partner UZH has recently developed the iCOBRA package for method benchmarking, which provides both a web application and R-package for reproducible and extensible benchmark experiment development.

Finally, RADIANT has supported a large number of courses, scientific meetings, developer workshops and partner-exchange visits to enhance training in computational biology, scientific software development, statistical data analysis and reproducible research. As well as training a large number of young scientists, materials from these activities are freely available online and provide a valuable resource for the community.

4.1.2 Summary description of project context and main objectives

The project concerned the development and rapid dissemination of computational tools for the analysis of data from a range of popular high-throughput sequencing (HTS) technologies. Work packages (WPs) were organised into three themes: (1) Analysis: Computationally efficient methods to optimally estimate biological quantities of interest from HTS data by controlling for stochastic error rates and by detecting and adjusting for systematic biases. (2) Integration: Helping us to understand the data at hand in the light of what is already known (e.g. from many other large-scale experiments), to achieve a systems-level understanding. (3) Dissemination: Making the progress made by individual researchers available to others by facilitating software method publication, distribution, benchmarking and training alongside theory publication.

The Analysis theme started with the fundamental task of mapping the raw read data without discarding informative biological signal (WP1). Other WPs in this theme addressed important open problems in the analysis of major classes of HTS experiments: genomic (WP2), epigenomic (WP4), transcriptomic (WP3, WP5) and interactome (WP3) data. In each case, novel methods for optimally extracting biological signal and comparing signal across experiments were developed and distributed through user-friendly open source software packages.

The Integration theme built on the methods from the Analysis theme, moving from the isolated analysis of a particular class of data towards systems-level modelling of biological function. In WP6, methods were developed for the combined analysis of data from Genomic, Epigenomic and Transcriptomic experiments. Bringing together these different but highly interrelated views greatly increases the power of association studies to reveal important regulatory interactions. In WP7, models of transcriptional and post-transcriptional regulation of gene expression were developed by combining results from different functional assays including RNA-Seq, smallRNA-Seq and ChIP-Seq data from different individuals or collected in time course experiments. In WP8, a visualisation tool was developed for an improved interpretation of data by displaying results of the analyses developed in the other WPs.

In the Dissemination theme our aim was to introduce a framework to encourage and support the publication of software alongside method development. In WP9, we supported developers in making their Bioconductor packages interoperable and consistent and to provide pathways through standard analyses. In WP10, we developed public benchmarking resources to help users and developers identify benchmark datasets, compare different related methods and optimise their experimental design and analysis strategies. In WP11, we developed training materials to support a comprehensive training programme for HTS data analysis.

All WPs together ensure that leading-edge statistical tools, including but not restricted to those developed in this project, are brought rapidly into the mainstream of biomedical research practice worldwide.

4.1.3 Description of main S & T results/foregrounds

Below we describe the main results for each work-package:

WP1: Improved read mapping

The aim of WP1 was to improve results of high throughput mapping algorithms in terms of resolving ambiguous mappings and identification of erroneous uniquely mapped reads to be able to get more reliable mapping results. In course of the evaluation, it became obvious that one of the more fundamental challenges in mapping – foremost when analysing small variants, but also in other cases – is the composition of the genome itself. Regions appearing multiple times in the genomic reference sequence (including but not limited to known genomic repeats) can make it tedious (or sometimes impossible) to assign a read to a single genomic position. Modern mapping algorithms take this into account and try to alleviate it by enabling users to set alignment quality limits. The downside is that high quality settings will reduce the amount of mapped reads in general, not just in ambiguous genomic regions. A "uniqueness background" of a reference sequence is therefore an essential basis for any kind of mapping quality approval and can already lead to a specific distinction between correct and erroneous mappings. This

can be crucial especially for clinical applications of HTS, where reliable rather than exhaustive results are of essential significance. Another issue with available mapping algorithms is that quality scores are not always easily comparable as they employ different scoring ranges. A more general approach to assess the quality of any mapping result would thus be helpful for comparing results from different mapping algorithms for casual users.

Uniqueness background modelling. Partner Genomatix have created a model to describe the uniqueness background for a reference sequence by “shortest unique subsequences” (SUS) for every position of the reference sequence. Based on this model we have developed an algorithm which is capable to rapidly identify the SUS up to a given length for every single position of a given reference sequence. The resulting SUS dataset can then be used for judging the correctness of an alignment. For the application of the methodology we implemented software (MS3, D1.1) which takes a BAM file and the generated shortest unique subsequence library as an input and assigns each alignment to one of three classes of decreasing stringency. Tests with artificially generated sequences (MS2) have shown that the first class will only contain alignments at correct positions with a very high specificity. Users can apply the software independently from the applied mapping algorithm and any differing scoring system to restrict the output to the most reliable set of results and thus get improved read alignments for any type of HTS application. The ultimate outcome is better individual read alignments. The independence of any scores should make it easily applicable even for casual or inexperienced users. In addition to the two algorithms outlined above we provide a "uniqueness" BAM file including coordinates and lengths of the SUS of the human genome hg19. This file can be uploaded in many genome browsers to visualize the uniqueness background of the human genome. It can also be used with any other software that can handle BAM files, e.g. to overlap or associate regions with the uniqueness background.

Improved alignment of bisulfite sequencing. A method was developed provide a fast, yet robust mapping algorithm for the identification of methylated patterns in DNA from bisulfite sequencing experiments and implemented in the **Genomatix Mining Station**. Our method uses a fast index-based algorithm that uses short unique subwords contained in the target genome or sequence based on the model described in the report for period 1. In addition to mapping the reads and providing the best positions in the reference, the software can also provide the mapping quality for second-best matches within a single run of the program, enabling users to perform a more stringent analysis in by only using well-defined mappings without the need to run a second mapping.

Bisulfite treatment converts cytosine ('C') residues in DNA to thymine ('T'), but leaves 5-methylcytosine residues unaffected. For the mapping task this leads to the complication that all 'T's in a read can either be a 'T' or could have been a 'C' originally. The sequence 'ATCGTTCGA' could match the references 'ATCGTTCGA', 'ACCGTTCGA', 'ATCGCTCGA', 'ATCGTCCGA', etc.

Common methods therefore convert reads to all possible combinations and then try to find the best match to the reference. This effectively reduces the four letter code representation of the DNA to a three letter code, leading to higher ambiguity and hence to a large number of copy hits. The advantage of using a suffix tree is that no conversion of bisulfite sequencing reads is necessary. We can simply walk through the tree of all short unique subwords to scan for cytosine/thymine conversions. If no unique subword is available from a 'thymine' node of the tree, one can assume that this was originally 'cytosine' and continue to search for matching words. The software can use read data from common sequencing platforms and produces positional file in standard formats (BED / BAM). In addition a methylation summary output file is generated providing statistical data for all positions in the reference where a cytosine nucleotide is covered by at least one bisulfite sequencing read.

The software has been developed in the C programming language with support for threads to make extensive use of multicore processors. It can be adapted to work in virtual machine environments of cloud computing platforms. We provide a command line executable for the Debian Linux platform; executable code for other platforms can be made available on request. We also provide index data for human and mouse reference genomes. Users can either use the software as a stand-alone mapper for bisulfite read data or integrate it into analysis pipelines

where necessary. Partner Genomatic has integrated the method into their commercial software solutions. It is used in our Genomatix Mining Station product and will be used in our extended Genomatix Genome Analyzer product within Q1 2016.

WP2: Identifying complex genomic variants

The main purpose of WP2 is the development of new methods to identify genomic variations from HTS data, with a particular focus on single nucleotide variations (SNV), copy number variations (CNV) and structural variants (SV). When the original case was written we focussed on the dominant short-read technologies available at that time but during the project the throughput and cost of long-read sequencing has improved and it seems likely that long-read technologies such as Oxford Nanopore's MinION will become a useful tool for detecting complex genomic variants which are very challenging for short-read technologies. We therefore decided to focus our own software development efforts on the processing of long reads produced by these technologies which hold greater promise than paired-end short reads for SV detection.

Detection of SNV and CNV from short DNA fragments. This task focuses on detection of variations from short DNA fragments. Partner ARMINES have developed, implemented and publicly released a software package for the detection of SNV from colour-space HTS data. The software, called COBALT, is the object of D2.1 which was delivered by the end of Month 18, as set out in the DOW. The method is based on an original probabilistic graphical model to model the particular structure of read generation in colour space. Thanks to the use of efficient online algorithms, it is about 20 times faster than existing methods and more accurate in SNV detection.

Detection of SV from paired end reads and long reads. We initially focussed on SV detection from paired end reads. We have identified an imaging artefact in the Illumina HiSeq equipment afflicting paired-end sequencing data where the two ends of fragments are inappropriately paired, leading to errors in SV detection from discordant read pairs. We have developed a method to diagnose instances where this has occurred and, if desired, to repair the affected reads. Members of the consortium collaborated with leaders in the field, including the ICGI consortium (UCAM) and Jan Korbel's team, developer of the DELLY software for SV detection (EMBL).
As new technologies producing longer reads have started to emerge (in particular Oxford Nanopore), we decided to focus our own software development efforts on the processing of long reads produced by these technologies which hold greater promise than paired-end short reads for SV detection. These efforts led to an R package called **IONiseR** (D2.2, see Part B2) for processing and quality assessment of long reads produced by Oxford Nanopore's MinION sequencer.

WP3: Differential Analysis of Count Data

For many assay type in high-throughput sequencing, standard analysis methods include the summarisation of the data by counting the number of sequencing reads that map to regions of interest (such as genes, exons, binding areas etc.) and then applying statistical analysis methods to this count table. WP3 is concerned with methods to obtain count data and to perform statistical inference on them.

Peaks, regions and counting schemes. The **htseq-count** method, developed at EMBL previously, continues to enjoy widespread use in the community (more than 200 unique visitors per day to the HTSeq web page) and an independent analysis by Fonseca et al. (2014) showed htseq-count to be superior to competing approaches. We released a new version of **HTSeq** and htseq-count, now with improved support for paired-end sequencing data (Anders et al, 2015, cited 281 times on Scopus). While the htseq-count script is suitable for standard RNA-Seq problems, the full HTSeq programming framework can be used by bioinformaticians wishing to develop custom counting tools for special cases or other assay types. A new tutorial-style chapter to the on-line documentation now explains in detail how to help new users leverage HTSeq for their needs.

In RNA-Seq data, counting bins are defined by gene annotation. In other areas, a good binning strategy has to be developed. For example, for HiC data, one may count read pairs by restriction fragment, by topological domain, or by some wider partitioning of the chromosome, and so get data suitable for different kinds of inference. We have explored this aspect in Pekowska et al. (2014). Related work studying this aspect for other assay types is ongoing.

GLM/ANODEV modelling. EMBL developed and released **DESeq2** (Love et al., 2014), a successor to their widely used DESeq tool for differential expression analysis for RNA-Seq data. UZH developed new features for edgeR (Zhou et al., 2014) another widely used tool for this task.

For DESeq2, we have combined the approach of generalised linear models (GLMs) with techniques from empirical-Bayes statistics to not only improve sensitivity and specificity of the method, but also offer conceptually new approaches to the analysis of HTS count data. For example, an important practical difficulty in RNA-Seq data analysis is that, due to the dependence of noise on count level, expression fold-changes cannot easily be compared between weakly and strongly expressed genes. DESeq2's conceptual advances, such as its shrinkage estimation of fold-changes and its new thresholded hypothesis tests, overcome this problem. DESeq2 is one of the most widely used tools in genomics with 228 citations to Love et al. (2014) already in Scopus. It is important that such methods are robust against the presence of outliers in the data, to avoid erroneous conclusions. This aspect was studied by both the EMBL team in the context of the development of DESeq2, and by the UZH team in the context of their continued development of **edgeR**. The latter work is described in the publication by Zhou X. et al. (2014), the former within Love et al. (2014).

Adaptation to novel assays. As we had hoped, the GLM/ANODEV methodology described above turned out to be suitable for other assays besides high-throughput sequencing. In fact, for several new assay types the adaption turned out to be more straightforward than anticipated, and several other researchers and users of DESeq2 succeeded in doing so merely guided by existing documentation. A special highlight here is that the very first CRISPR/Cas9 screen, reported in Nature (Zhou Y. et al., 2014), was analysed with DESeq2, showing the value and relative simplicity of adapting this RADIANT-developed method to a groundbreaking new technology.
We concentrated our own efforts on a class of novel assays whose importance for functional genomics research has become apparent in the last few years. These are high-throughput chromosome conformation capture assays, specifically 4C, HiC and ChIA-PET. We developed the Bioconductor package **FourCSeq**, accompanied by a research paper in Bioinformatics, (Klein et al., 2015), and applied it in research reported in Nature (Ghavi-Helm et al., 2014), and presented further results on analysis of HiC data (Pekowska et al., 2014).
We also worked on analysis of RNA-Seq data from single-cell sequencing experiments, and described the first comprehensive method to properly account for the strong technical noise inherent to this assay type (Brennecke et al., 2013),. We finally mention the work of Gupta et al. (2014), where we demonstrated the use of DESeq2-related techniques on CLIP-Seq data.

Variance stabilising transformation. Many standard techniques in statistics and exploratory data analysis require data to be approximately homoskedastic, i.e., the variance of the data should be independent from the mean. Due to Poisson noise, this requirement is not met by count data, which precludes, for example, the application to RNA-Seq data of many techniques commonly used in microarray data analysis. The purpose of variance-stabilising transformations (VSTs) is to transform data in a manner that makes them suitable for use with such techniques. At EMBL, in the development of DESeq2, we have found that the empirical-shrinkage techniques we developed can also be used for this purpose and hence included functionality to this extent, which we termed the "regularized logarithm" ("rlog") transform. Assessing the performance of this transformation for purposes of exploratory data analysis is ongoing work. Similarly, at UZH, we have explored the suitability of the voom method (Law et al., 2014) for such tasks.

Time series differential expression for read counts. UNIMAN and USFD collaborated to develop a new algorithm for identifying differentially expressed genes from time-course data. It is implemented with an existing software package for Gaussian Process, known as GPy,

developed at USFD. The new algorithm is implemented as a dedicated sub-module, a kernel function, and integrated into the GPy software package. This sub-module, to become publicly available soon, is currently used to identify at which time(s) a gene becomes differentially expressed under different conditions. Due to its generality, it can potentially be applied to identify the time of perturbation in any typical two-sample problems. As the new algorithm is an extension of the existing framework, it can make use of all the implemented noise models, e.g., Gaussian, Gamma, Poisson, Student-t, etc., and take advantage of sophisticated approximate inference algorithms already developed in GPy. These features differentiate this algorithm from current stand-alone two-sample data analysis algorithms, and makes two-sample data analysis with sophisticated noise models possible. The method has also been developed as a stand-alone R package DEtime by UNIMAN (available on GitHub) and will be submitted to Bioconductor. In related work, we have also developed a hierarchical approach to modelling gene expression time course data within GPy which allows for better modelling of biological replicates with uneven sampling of time points and missing data (Hensman et al., 2013).

WP4: DNA Methylation Data Analysis

The main objectives of WP4 are in the development of computational tools for the integrative analysis of DNA methylation data. IIT have developed three companion R packages providing key functionalities in line with the expected WP objectives: **methylPipe**, **compEpiTools** and **ListerEtAlBSseq**. Briefly: (i) *methylPipe* includes multiple classes, methods and functions to deal with base-resolution DNA methylation data, supporting 5-methyl cytosines and 5-hydroxy-mC in either the CpG or non-CpG sequence context; (ii) *compEpiTools* provides functions and methods for the integrative analysis of DNA methylation and other epigenomics data; (iii) *ListerEtAlBSseq* provides two complete base-resolution methylomes, derived from the first two whole-genome DNA methylomes in human (Lister R et al, Nature 2009), including data from two well studied cell lines (H1 embryonic stem cells and the IMR90 fetal lung fibroblasts).

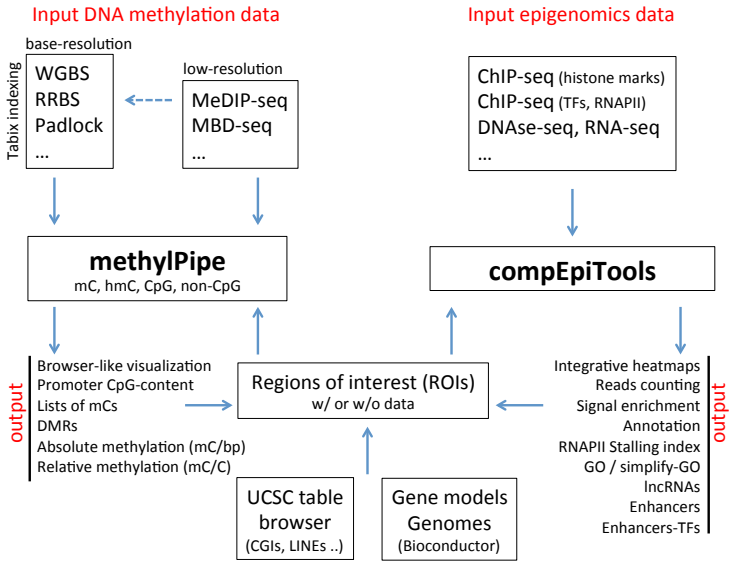

Figure 1: Diagram describing input and output for the methylPipe and compEpiTools packages.

These three R packages were included in Bioconductor since December 2014 and have more than 6300 downloads so far. A manuscript describing these tools and illustrating how they could be combined to perform an integrative analysis of various epigenomics and regulatory omics data was published (Kishore et al. 2015) with more than 2100 accesses on the BMC web pages since Sep 29th 2015.

Automatic report generation. The visualization routines were extended well beyond the initial emphasis given on this point. They were considered mature and self-standing and we decided to set them as the basis of a complimentary while independent project, the compEpiTools package. This is heavily based on the development of a comprehensive heatmap-based system for effortlessly integrating and visualizing heterogeneous omics data. Importantly, the compEpiTools package is fully compliant with objects and results obtained with the methylPipe package, emphasizing the integration between DNA methylation and other omics data types.

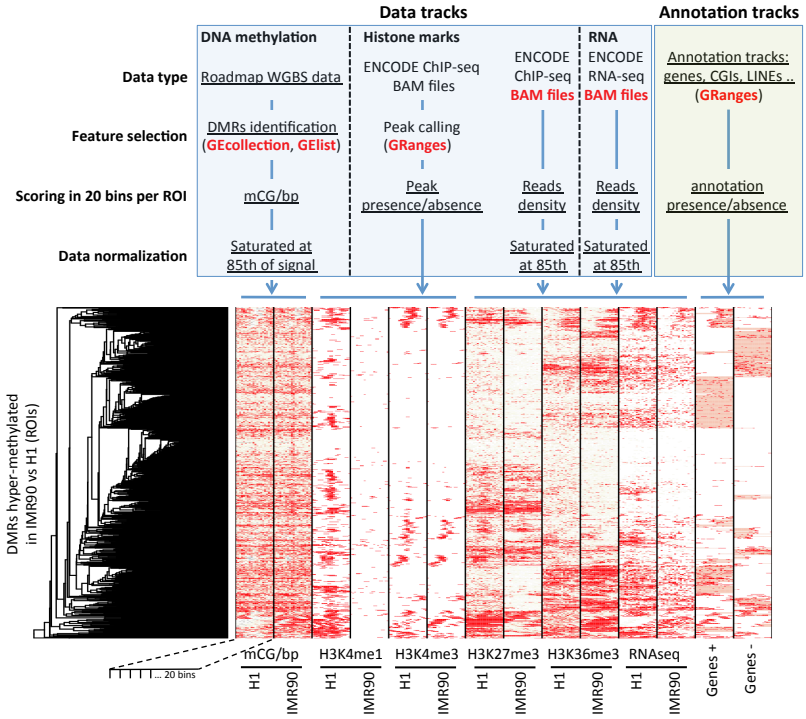

Figure 2 An integrative heatmap generated with the compEpiTools package.

WP5 Transcript-level Expression Estimation and Comparison

RNA-Seq data provide the opportunity to identify different transcript isoforms generated from a gene locus and to distinguish between alleles (in a diploid organism, typically two). Across different experimental conditions, tissue types etc., they allow us to ask whether isoform usage or allele-specific expression is differential, i.e. shows evidence of regulation. A complication is raised by the fact that current HTS produces short reads, which are generally not long enough to directly identify the isoform. Probabilistic models have been introduced to estimate transcript expression levels for genes with multiple alternative isoforms and/or allelic variants from RNA-Seq data. Bayesian methods provide an attractive approach as they can be used to quantify the uncertainty and covariation in expression estimates from closely related transcripts. However, current Bayesian approaches rely on Gibbs samplers to generate samples from the posterior distribution of transcript expression levels, which can be computationally intensive. In this objective we have therefore developed efficient algorithms to make advanced transcript-level modelling more practical. We also investigated method for discovering novel transcript structures and investigated differential transcript usage methods across conditions.

An efficient algorithm for identifying unannotated transcripts. ARMINES have introduced a new technique called **FlipFlop** (Bernard et al., 2014) which can efficiently tackle the sparse estimation problem on the full set of candidate isoforms by using network flow optimisation. The technique removes the need of a pre-selection step and leads to better isoform identification while keeping a low computational cost. Experiments with synthetic and real RNA-Seq data confirm that our approach is more accurate than alternative methods and one of the fastest available. FlipFlop has been implemented as a Bioconductor package.

Efficient approximate inference algorithms for transcript expression inference. UNIMAN and USFD have been working together to introduce a novel Variational Bayes (VB) inference scheme for the **BitSeq** transcript expression estimation and DE scoring algorithm (Hensman et al., 2015; Papastamoulis et al., 2014). This provides a speed-up over the existing MCMC and published VB schemes. This approach allows the marginal likelihood to be computed more accurately than standard VB-EM algorithms and this result can be used to score alternative transcriptome definitions and thereby identify incorrect gene models. The VB algorithm is

implemented in the BitSeq Bioconductor package. We provide a comparative study against seven popular alternative methods and we demonstrate that our new algorithm provides excellent accuracy and inter-replicate consistency while remaining competitive in computation time. Hensman et al. (2015) was the most-read article in the journal Bioinformatics in December 2015. We have developed the code as a stand-alone package in GitHub and also incorporated it in the Bioconductor version of BitSeq (D5.2).

Count-base gene-level differential splicing analysis. UZH have compared current state-of-the-art methods for differential transcript usage as well as suggesting improvements to commonly used workflows (Soneson et al. 2016). The performance of representative workflows was assessed using synthetic data and we explored the effect of using non-standard counting bin definitions as input to a state-of-the-art inference engine (DEXSeq, developed by EMBL). Although the canonical counting provided the best results overall, several non-canonical approaches were as good or better in specific aspects and most counting approaches outperformed the evaluated event- and assembly-based methods. It was shown that an incomplete annotation catalogue can have a detrimental effect on the ability to detect differential transcript usage in transcriptomes with few isoforms per gene and that isoform-level pre-filtering can considerably improve false discovery rate (FDR) control. Count-based methods generally perform well in detection of differential transcript usage. Controlling the FDR at the imposed threshold is difficult, mainly in complex organisms, but can be improved by pre-filtering of the annotation catalogue.
Current follow-up work at UZH is on the development of a new package for Differential Transcript Usage using a Dirichlet-Multinomial model to capture changes in transcript proportions while also modelling biological replicate variance. In addition, EMBL have continued developing DEXSeq, for exon-level differential testing, and have adopting statistical methods from DESeq2 in the most recent versions.
UZH have also assessed the impact of transcript-level inference methods on differential expression studies and shown that aggregation of transcript-level results to the gene level typically improves performance of differential expression tests over transcript-level tests (Soneson et al. 2015).

WP6 Genome, Epigenome and Transcriptome

The main purpose of WP6 is for the development of integrative genomic data analysis methods; that is, development of models and methods for more than one (large-scale) data type simultaneously. This is a challenging and broad area of research, since there is a large space of associations between data types that researchers are interested in and especially in clinical/medical settings there may be significant sources of noise and bias limiting data quality and availability. Our work package has made significant progress in integrating genetic information (e.g. copy number) into analyses of DNA methylation and ChIP-seq data. We have developed and distributed software (described below) through the Bioconductor project and are actively engaged in further integrative data analyses together with collaborators (e.g. DNA methylation and gene expression). Robinson (UZH) and Pelizzola (IIT) edited a special issue in Frontiers in Genetics focused on the nascent field of computational epigenomics, including an editorial and one article each from the UZH and IIT groups (de Pretis et al. 2014a; Robinson et al. 2015).

Integrative pre-processing for quantitative sequencing dataset. Partner UZH have developed methods for the integration of copy number variation and other assay-specific factors into methods for ChIP-seq (differential) and affinity-based DNA methylation (absolute and differential) analyses. Software has been made available in the Repitools R/Bioconductor package (Riebler et al. 2014).

Integrative analysis – epigenome state. The INSPEcT R/Bioconductor package was developed by IIT for studying the dynamics of transcriptional regulation (de Pretis et al. 2014b). They are now using this tool to study how these dynamics are associated with the Pol2 and epigenetic machinery following the activation of a master transcription factor (Myc).

Integrative analysis – transcription outcome. The modelling-based activities of this task were geared to associating expression and splicing outcomes with genetic or epigenetic factors. Hi-C data were used to take into account the 3D structure of the genome and Partner ARMINES developed a method and associated software to infer the 3D structure from these data (Ay et al., 2014) and to analyse the DNA 3D structure influences gene expression (Varoquaux et al., 2015). Using regression and supervised classification-based techniques, we showed that the direct hypermethylation of gene promoter regions often observed in cancer is unlikely to directly inhibit gene expression. Although there is no immediate plan to make a package, R software used to carry out these studies are publicly available as supplementary materials of the corresponding publications.

WP7 Statistical Methodologies for Systems Level Modelling

The objective of this work package for this period was to integrate different -omic datasets derived from HTS technologies in order to infer systems biology models of interaction, regulation and dynamics.

Learning models of transcriptional regulation by integrating multi-modal time-series data. UNIMAN, USFD and Genomatix collaborated on a study integrating pol-II ChIP-Seq data, pre-mRNA data from RNA-Seq intronic reads and mRNA-Seq data in a model of transcriptional activation (Honkela et al. 2015). We found significant pausing between transcription completing and mRNA production in a sub-set of genes and analysis of intronic reads in the RNA-Seq data suggesting a splicing-associated delay. Methods from WP5 (BitSeq) were used to aggregate transcript-level inference into gene level summaries for time course modelling and this was shown to improve modelling performance. The method was distributed as Matlab code (based on an older code-base) but is now being developed into an R/Bioconductor package building on methods in the tigre package.

Identification of microRNA regulated genes. microRNAs are small RNAs repressing target mRNAs' expression. We aimed at exploiting this property of microRNA in order to identify their target genes from gene expression profiles. In order to develop and validate a computational approach to microRNA target gene identification, we exploited a dataset derived from human retina samples consisting of RNA-seq and smallRNA-seq of 50 human donors (described below). Partner TIGEM built a bioinformatics pipeline for reverse engineering gene and microRNA regulatory networks from multiple gene expression profiles measured using RNA-sequencing technology. The main concept behind our computational approach is to exploit the inter-individual variability to reconstruct (reverse engineering) gene/microRNA co-expression networks. In such a network two entities (genes or microRNAs) are connected by an edge if their expression profiles across the sample are statistically dependent. This dependency can be estimated in several different ways. In our pipeline, we focused on pair-wise measures including Pearson correlation coefficient, Spearman correlation coefficient and mutual information, which are user-selectable.
Once a network is obtained, microRNA targets can in principle be identified by selecting those gene-microRNA pairs in the network with a negative SCC or PCC, as one can assume that a microRNA and its target should be negatively correlated. D7.2 describes the application of the pipeline to a dataset generated from sequencing the polyA RNA and smallRNA of 50 retina samples from post-mortem donors. We demonstrated that the pipeline yields an informative network when polyA RNAs are considered. However, we could not validate the microRNA-target discovery potential of our tool, since out of the 50 samples, only 16 yielded high-quality reads for the smallRNA-seq, hindering the application of our approach.

Identification of post-translation regulation. Through modulation of transcriptional regulation, a cell is able to tightly regulate protein function such as its activity, localisation and interaction with other molecules. Capturing this kind of regulatory interactions using only transcriptional data, such as gene expression profiles (GEPs), is considered challenging since GEPs are the resulting effect of multiple regulatory events. Here, we modified a computational strategy we previously developed for microarrays measurements, called Differential Multi-Information (DMI),

to infer post-translational modulators of a transcription factor from a compendium of gene expression profiles (GEPs). DMI is built on the hypothesis that the modulator of a TF (i.e. kinase/phosphatases), when expressed in the cell, will cause the TF target genes to be co-expressed. On the contrary, when the modulator is not expressed, the TF will be inactive resulting in a loss of co-regulation across its target genes. DMI detects the occurrence of changes in target gene co-regulation for each candidate modulator, using a measure called Multi-Information. We validated the DMI approach on a compendium of 5,372 GEPs showing its predictive ability in correctly identifying kinases regulating the activity of 14 different transcription factors.

Application to a cohort study. TIGEM generated two datasets of high quality transcriptome and miRNome data from 50 human retina samples. We used the Spearman Correlation Coefficient (SCC) based method to quantify the co-expression of all the gene-pairs across the 50 samples using the expression values in the RefT. The resulting gene network consists of 11,189 genes and 2,458,363 edges connecting gene-pairs genes, which corresponds to about 2.5% of all the possible gene-pairs. About one half of the genes (53%) had more than 100 connections. In order to assess the biological significance of the gene network, we used as a gold standard the functional and physical gene-gene interactions reported in the STRING database. We then sorted gene-pairs in the network according to their SCC values and checked whether each gene-pair is supported by STRING. We estimated the percentage of correct connections, had these been randomly guessed, to be equal to 0.02%. These results confirm that the gene network we inferred by exploiting inter-individual variability in gene expression contains biological relevant information.
Gene networks can be used to assess the function or tissue-specific expression of a gene via a guilty-by–association approach. This approach consists in assigning a function to a gene by checking whether there is a shared function among its ‘gene neighbours’, i.e. the set of genes connected to it in the network. We used the guilty-by-association approach to identify genes specifically expressed in photoreceptors cells.

WP8 Advanced Data Visualisation along Genomic Coordinates

The RADIANT Genome Browser is based, technically, on the Savant Genome Browser and on its plugin technology. The Savant Genome Browser (http://genomesavant.com/p/home/index) is a Java application, has been designed with the HTS datasets in mind, is open source, computationally performing and user friendly, with an extensive plugin framework which can be completed and developed.
This software is designed under the Model-View-Controller software engineering paradigm. That is, the code which governs input and presentation elements (User Interface and visualization, respectively) is isolated from the application logic (data reading, writing and manipulation). We adopted this concept for our development in the project.
A Radiant Framework has been organized and it is currently used by two RADIANT work package implementations (application plugins). The framework has two separate sub packages: the first collects a number of graphical widgets that can be used in combination to implement any descriptive statistics panel; the second implements a collection of tracks elements used for interval-based visualization.
Genomnia implemented the Bioconductor methylPipe software output visualization plugin (WP4), focused on the analysis of genome-wide methylation sequencing datasets. The second RADIANT application plugin reads and displays Hi-C datasets (WP2).

WP9 Helping Scientists to Publish and Use High Quality Statistical Software

The overall aim behind WP9 was two-fold: to help scientists – in the consortium and beyond – with the publication and continued maintenance of their software, and to provide prospective users with guidance on finding, using and efficiently combining packages, typically from multiple authors, in order to solve their scientific questions. This was a broad aim, which has required a multitude of efforts and resources. The work package’s specific aim was to leverage on-going efforts and to make a significant contribution by addressing bottlenecks. A developer-writer (A. Oleś) was employed through the project to produce reproducible research case studies of high-

throughput sequencing data analyses associated with their scientific publications.
The benefits of a functioning and vigorous ecosystem of statistical academic software are manifold: scientists who publish software gain impact and professional stature; users are enabled to employ advanced methodology that they typically do not have the time or expertise to implement themselves; small enterprises can focus on their core competencies and choose best algorithms from a large, pre-competitive corpus of academic expertise.
Successful relevant efforts already existed in the field of statistics for high-throughput sequencing (HTS) data. These include centralized software repositories such as Bioconductor[2], Sourceforge, CRAN, as well as indices of distributed resources, e.g. Bioinformatics Application Notes or the SEQanswers software portal wiki; fora, in which developers and users of methods meet, either online (e.g. SEQanswers forum or Bioconductor mailing list) or in the real world in advanced courses and workshops organised by EMBO, Wellcome Trust, ISCB, Bioconductor and others; and common infrastructure software projects, such as the infrastructure packages produced by the Bioconductor core team.

Helping users to solve scientific problems by applying and combining available HTS software.
Our contributions include a protocol paper (Anders et al., 2013) providing guidance on current best practices and presenting a state-of-the-art computational and statistical RNA-Seq differential expression analysis workflow based on R and Bioconductor software and, in particular, on two widely used tools, *DESeq* and *edgeR*. As an extension of work on WP3 (the follow-up version of *DESeq*, i.e. *DESeq2*), we have produced comprehensive documentation which both emphasize the building of complete analysis workflows, not only using a single package but the best available components. Specifically, we not only overhauled the existing technical documentation, but also authored a "Beginner's Vignette" (Love et al., 2014b, http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/beginner.pdf), which offers a gentle but comprehensive introduction to the topic of RNA-Seq data analysis. This and similar introductory documents are crucial to make modern biostatistics tools accessible to a broad community of researchers by helping newcomers getting started.
Along similar lines, the vignette of the *parathyroidSE* experiment data package on Bioconductor now describes the construction of the *SummarizedExperiment* object suitable for differential gene expression analysis "from scratch", i.e., from a collection of FASTQ files as obtained from the sequencing core facility, or downloaded from authors' publications (e.g. via EBI/ArrayExpress/SRA).
To reach an even broader audience and to make HTS analysis more accessible to users not proficient in R, an effort to develop an RNA-Seq workflow for the Illumina BaseSpace cloud computing environment is ongoing.
Regarding the topic of task views, i.e. overviews of the software offerings in particular areas, V.J. Carey already leads this effort in the *BiocViews* Bioconductor package. Moreover, user uptake and demand in this case seems to be modest; therefore we have decided to refocus our attention on problem-oriented documentation providing solutions to a selected set of specific scientific questions and guidance on performing end-to-end analyses, i.e., workflows and vignettes: computationally reproducible documents demonstrating complete analysis pipelines by calling a sequence of required tools embedded in an explanatory narrative. We have set up a 'Bioconductor channel' in the open-access, post-publication-peer-review journal *F1000Research* (http://f1000research.com/channels/bioconductor). The channel provides a forum for task-oriented workflows that each cover a solution to a current, important problem in genome-scale data analysis from end to end, invoking resources from several packages by different authors, often combining multiple `omics data types, and demonstrating integrative analysis and modelling techniques. The channel addresses an urgent, and previously poorly satisfied need for cross-package, task-centred workflows that address common data analytic workflows from end-to-end. This problem-oriented documentation is in contrast to tool- or method-oriented documentation that often accompanies particular pieces of software and addresses the question "What can I do with this particular package?" rather than the question that motivates the channel and the workflows it contains (and is currently accumulating): "How do I solve this particular problem?"

[2] http://bioconductor.org

Members of the consortium were also involved in training as described in the WP11 report. In 2013, 2014 and 2015 we co-organized and we provided lectures at CSAMA: Computational Statistics for Genome Biology, a one-week intensive course on current approaches in statistical and computational analysis of large-scale experiments in biology, held in Bressanone (Italy). The course focused on methods for downstream analysis of HTS experiments including DNA sequencing (variant calling), RNA sequencing (differential expression), QTL analysis, and epigenetics. Other teaching activities that fall under this task include:

- EMBO Practical Course Bioinformatics and Statistics for Large-Scale Data, Shenzen, China, 17-22 November 2013 (Wolfgang Huber and Simon Anders, EMBL) http://events.embo.org/13-large-scale-data/index.html
- RNA-Seq analysing using Bioconductor course, X-Meeting, 3-6 November 2013 (Wolfgang Huber, EMBL) http://x-meeting.com/rna-seq
- EMBL Advanced Course on High-throughput Sequencing, Heidelberg, 23-25 September 2013 (Simon Anders, EMBL)
- EMBO Practical Course on Analysis of High-Throughput Sequencing Data, Hinxton, 21-26 October 2013 (Simon Anders, EMBL)
- EMBO Practical Course on Analysis of High-Throughput Sequencing Data, Hinxton, October 2014 and October 2015 (Wolfgang Huber, EMBL).
- EMBL course Statistical Bioinformatics using R and Bioconductor, 13-16 October 2014 (Four full days) http://www.embl.de/training/events/2014/RPD14-01/index.html
- Basic R, data handling and graphics, ZMBH Heidelberg for the SFB 1036, CELLULAR SURVEILLANCE AND DAMAGE RESPONSE, 27-29 May 2015 (Three full days) (Bernd Klaus, EMBL).
- Workshop (1.5 hrs) on RNA-Seq analysis at the summer school "To see the (Black) Forest for the trees: Black Forest Summer School on NGS data for phylogenetics" - 16 September 2015, http://plantco.de/BFSS2015/ (Bernd Klaus, EMBL).
- EMBL course Statistical Bioinformatics using R and Bioconductor, 12-16 October 2015 (Five full days) (Bernd Klaus, EMBL) http://www.embl.de/training/events/2015/RPD15-01/index.html.

Helping developers to produce interoperable, best practice software. We have continued to collaborate with the Bioconductor core team and have participated in reviews of new packages to help ensure that the required software quality standards and good coding practices are met, namely: comprehensive documentation, modularity, portability, interoperability, unit testing, and computational performance.

Help with package creation: We provide consultancy and assist members of the network and beyond on developing R packages. We continue to engage with scientists from in- and outside the consortium to help deliver their methods as best-practice, interoperable and easy-to-use software, in particular in the form of Bioconductor packages.

Package interoperability: Data exchange between popular Bioconductor HTS data analysis packages, such as *edgeR*, *DESeq2*, *baySeq* or *DiffBind*, is hindered because of different data containers used for storing read counts and analysis results. We aim to address opportunities for improved interoperability between these packages by providing a converter package facilitating data exchange across various differential gene expression data formats in Bioconductor. Our initial implementation which contains a converter between DESeqDataSet (*DESeq2*) and DGEList (*edgeR*) objects has been released on GitHub[3] in the R package *DEFormats*.

Workflow channel: as mentioned above, we have set up a 'Bioconductor channel' in the open-access journal *F1000Research* (http://f1000research.com/channels/bioconductor). Besides the task of user support, described above, it also helps developers and prospective authors, by providing an incentive to actually produce and author such workflows. In particular, it provides a highly visible, peer-reviewed, citable, PubMed-indexed dissemination platform for such documents.

Reproducible research.

In this work package scientists of the consortium were assisted by a developer-writer to

[3] https://github.com/aoles/DEFormats

produce reproducible research case studies of high-throughput sequencing data analyses associated with their scientific publications. The most important examples of the case studies and results are highlighted below.

Standard formatting style: To lower the entry barrier and to ease the process of publishing appealing documentation even by non-expert package contributors, we have in Period 2 further developed and improved the R package *BiocStyle,* which provides a standard formatting style for Bioconductor literate programming vignettes and other types of documentation. It features a consistent layout for PDF and HTML documents, with entry points being either Sweave- and knitr-type .Rnw markup source files, or the newer markdown .Rmd files. The user response to the package has been exceedingly positive. Since its introduction in 2013 the number of packages using *BiocStyle* has doubled every half a year simultaneously to each new Bioconductor release. In the current Bioconductor version 3.2, 262 packages depend on *BiocStyle*, which is almost one fifth of all Bioconductor software and experiment data packages. We have consulted a professional graphic designer who specialises in layouting of technical literature. She has provided us with very valuable advice on the layout of document elements including pages, paragraphs, figures, code blocks, headings, etc. We are currently implementing these design ideas into the software (using funding from other sources).

The *BiocStyle* package can also be used to produce appealing "reproducible research" documents, i.e. documents using the "literate programming" paradigm, for the communication and dissemination of analytical workflows. It has been used, for instance, for the Bioconductor workflows (http://www.bioconductor.org/help/workflows) and for teaching materials (http://www.bioconductor.org/help/course-materials/).

Supplementary materials facilitating reproducibility: Recent discussions within the life sciences community have highlighted the importance of means to document the computational methods used in research in a manner that guarantees completeness, accuracy and reproducibility. Bioconductor data packages that accompany scientific publications are a very suitable way to do so, as we could demonstrate in several cases where Radiant members contributed their bioinformatics expertise to biology projects:

The *Hiiragi2013* Bioconductor experiment data package contains a complete executable transcript (vignette) of the statistical analysis presented in the paper "Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages" (Ohnishi et al., Nature Cell Biology, 2014). It allows reproducing all results in the paper including figures, tables, statistical tests, etc.

The *PGPC* experiment data package that accompanies the paper by Breinig et al. (Molecular Systems Biology, 2015) encapsulates the analysis pipeline and the data files in a similar way as the *Hiiragi2013* package in form of a self-contained Bioconductor package. It also features a vignette, which guides through the analysis of the chemical-genetic interaction screen in isogenic HCT116 cell lines and provides code to generate all results and figures from the manuscript. Apart from the *PGPC* Bioconductor package, there is also an on-line PGPCviewer tool[4] hosted at EMBL Heidelberg and publicly available. It was built using Shiny. PGPCviewer offers users the possibility to search for compounds of interest, investigate respective phenotypic chemical-genetic interaction spectra and directly inspect phenotypic effects of drugs on cells by providing all microscopic images obtained in the experiment in an easy to browse format.

The *DESeq2paper* R package[5] contains literate programming documents for rendering all figures and tables of Love et al. (Genome Biology, 2014) together with data objects for the experiments mentioned in the paper, and code for aligning reads as well as for benchmarking.

The supplementary files of the manuscript by Brennecke et al. (Nature Immunology, 2015) contain a vignette, which reproduces each statistic and figure from the manuscript. The associated R package has been submitted to the Bioconductor repository and is currently under revision.

Our approach to reproducible research has thus followed the advice by Prof. Jeff Leek "Instead of research on reproducibility, just do reproducible research" (http://simplystatistics.org/?p=4563).

Interactive Data Exploration Tools: Another approach was used by partner UZH (Zhou et al.

[4] http://dedomena.embl.de/PGPC/

[5] http://www-huber.embl.de/DESeq2paper

2014), who provide a website[6] containing supplementary data together with the R-code for rerunning the simulations and analysis. Additionally, they offer an on-line browser[7] for exploring RNA-seq count datasets simulation, by leveraging the R/Shiny platform in a similar fashion as the PGPCviewer by Breinig et al.
Also Klimmeck et al. (2014) used Shiny to build an on-line browser[8] as an interactive supplement to their paper. Their browser provides public access to the data through an easy and friendly interface allowing data exploration of the transcriptome and proteome of the LS-K cells and LS+K cell populations of mice without any need for programming skills.
Interactive Software Documentation with Jupyter notebooks: The Jupyter notebook[9] is an interactive environment for code evaluation, exploratory data analysis and reporting originally developed for the Python language. We facilitated its use in bioinformatics analysis working with R and Bioconductor. Docker[10] provides a system of images and containers for running applications in a Linux-like environment without the need of full virtualization. The base Docker image[11] for our IPython toolset can be found in the registry of Docker Hub[12] under the repository name: **vladkim/ipynb**. Essentially, this base image contains all the dependencies required for running IPython notebooks together with the R kernel[13]. We have described our strategy for distribution of Jupyter notebooks with Docker in the Deliverable document 9.1 and a demonstration of feasibility of such a workflow is also available as a video[14]. Another important contribution is our package RdocsJupyter[15] that automates the conversion of a static vignette to a Jupyter notebook and subsequently builds a Docker image that can be pushed to Docker Hub and delivered to other users. See D9.1 (Report on Reproducible Research) for further details.

WP10 Benchmarking Statistical Methods and Experimental Protocols

In this work package we proposed to collect publicly available data and data available from collaborating experimental groups to create a benchmarking resource for HTS analysis methods. These benchmark were to be made available as a desktop and web-based application in order to help researchers make informed choices over which tools and experimental protocols to use for a particular application.
Benchmarking sequencing data applications often involves either differential analysis of mapped read summaries across replicated conditions or classification of samples characterised by high-dimensional vectors of summaries as features. The iCOBRA package (Soneson & Robinson, 2015) has been developed at UZH to provide a flexible general-purpose web-based application and accompanying R package to evaluate, compare and visualize the performance of methods for estimation or classification when ground truth is available.
Benchmarking ChIP-seq peak calling and region identification/enrichment algorithms have proved problematic as there has been no extensive validation of sites using an alternative experimental technique to provide an independent ground truth. Most publications only validate a very small number of sites, making collecting a comprehensive dataset difficult. UCAM have investigated and benchmarked statistical models to fit the distribution of read counts observed in ChIP-seq data (Cairns et al. 2014).
More generally, we have established a website (http://radiant-project.eu/Output/Benchmarking.html) to serve as a catalogue of data that may be useful for benchmarking. As an inclusion criterion, we have required in each case that the primary data were generated on a next-generation sequencing platform, most frequently the Illumina HiSeq, and must be available for download (although they require permission from data access committees). In addition to this there are data generated on an orthogonal platform from the same biological samples. This may take the form of microarray data, whole sample sequencing

[6] http://imlspenticton.uzh.ch/robinson_lab/edgeR_robust/
[7] http://imlspenticton.uzh.ch:3838/robust/
[8] http://vega.embl.de/LSK/
[9] https://jupyter.org/
[10] https://www.docker.com
[11] https://hub.docker.com/r/vladkim/ipynb/
[12] https://hub.docker.com/
[13] https://github.com/IRkernel/IRkernel
[14] https://www.youtube.com/watch?v=sxn-sixRVtYsK
[15] https://github.com/vladchimescu/RdocsJupyter

using an alternative NGS platform, or targeted sequencing of specific genomic loci. Each of these verification approaches has its own limitations - e.g. alternate NGS approaches are potentially susceptible to the same biases that produced incorrect calls in the original data, while microarrays and targeted approaches are limited in the number of loci they can interrogate. Targeted verification approaches are particularly problematic as the location of the target regions was, in the majority of cases, driven by calls made from the original sequencing data. Thus, it is unlikely that a novel call generated by the method being benchmarked will have any corresponding verification data. This is compounded by the fact that the verification status is often published simply as a binary field in a table of calls, and those that fail verification are often silently excluded. Nonetheless, these relatively short lists of verified variants are often used to provide some measure of the accuracy of variant calling software. The intention is for this site to make it easier to locate such data, particularly in fields such as cancer genomics, and to bring together complementary data that were run in different experiments and may otherwise be missed. Each dataset is accompanied by a brief description, where we have tried to detail the types of samples studied and the technologies used. Where appropriate we have attempted to highlight the locations of relevant information (e.g. lists of validated variant calls) in a publication and in some cases, simplified forms of such data are provided directly. Any ambiguities, such as whether sites that failed verification are still included in lists of potential variants, are also discussed.

4.1.4 Description of potential impact (including socio-economic impact and wider societal implications of project) and main dissemination activities and exploitation of results 10 pages max

The RADIANT project provides significant improvements in computational tools for the analysis of high-throughput sequencing (HTS) data and these tools are already having a significant impact on scientific research. These are the tools that cutting-edge life-sciences research needs right now, so enabling progress in biomedical research. Table A2 lists an impressive range of software packages developed or improved through this project. The DESeq2 tool is one of the most popular tools in current genomic research (34210 downloads this year; Love *et al.* Genome Biology 2014 paper already cited 228 times in Scopus) with diverse applications, e.g. to the first CRISPR/Cas9 screen, to single-cell transcriptomics data and to nascent CLIP-Seq data. RADIANT has been at the forefront of developments in transcriptomics with improvements to popular packages edgeR, HTSeq and BitSeq, but has also developed methods for newer HTS technologies such as HiC (FourCSeq), methyl-seq (methylPipe) and MinION long-read data (IONideR) which are having increasing impact on biomedical research as exemplified by our collaborative work (e.g. Ghavi-Helm et al. *Nature* 2014; Ay et al. *Genome Research* 2014; Sabo et al. *Nature* 2014). ***With the growing application of genomics in medicine and agriculture then the impact of many of these popular, powerful and robust statistical tools is expected to be widespread, especially given the popularity of the R/Biocoductor project where most of our new tools are now available.***

A major theme of RADIANT was to improve the critical software development infrastructure required for successful and reproducible computational research (as detailed extensively in Section 4.1.3, WP9). Our development of the BiocStyle package provides standard formatting style for Bioconductor vignettes and documentation which will greatly improve the Bioconductor user-experience. Through the 'Bioconductor channel' of the *F1000Research* journal and in the development of converter packages we provide task-centred support for genomic data analysis which is often much more useful for the practitioner than standard method-orientated documentation. Another key piece of infrastructure is a uniform framework for method benchmarking and the recent iCOBRA provides a useful framework which is available both as a web-tool and R package for HTS data analysis method benchmarking. ***With these developments in scientific software publishing and method benchmarking we are providing solutions to the reproducibility crisis in the life sciences and demonstrating best practices.***

Within WP11 we supported a range of superb activities for dissemination of computational methodology and for training the next generation of computational scientists. Table A2 lists a

number of high-quality courses, developer workshops and scientific meetings supported by RADIANT which have informed large numbers of academic and industry-based scientists about leading-edge computational developments. ***Our provision of superb training and training resources will continue to support the development of the next generation of computational scientists beyond these specific activities and help close the skills gap in this key discipline.***

All of these developments are contributing to overall well-being of society at large by providing crucial contribution to sciences that underlie progress in medicine.

4.2.5. Website and relevant contact details

The RADIANT project website is at http://radiant-project.eu

For further information about the project, please contact Professor Magnus Rattray (Scientific Coordinator) magnus.rattray@manchester.ac.uk

RADIANT Partner Contact Details

No.	Institute	Short name	Location Country	Principal Investigator Contact e-mail
1	University of Manchester	UNIMAN	Manchester UK	Magnus Rattray magnus.rattray@manchester.ac.uk
2	Genomatix	GENOMATIX	Munich Germany	Korbinian Grote grote@genomatix.de
3	Genomnia	GENOMNIA	Milan Italy	Alessandro Guffanti alessandro.guffanti@genomnia.co
4	European Molecular Biology Laboratory	EMBL	Heidelberg Germany	Wolfgang Huber whuber@embl.de
5	Telethon Institute of Genetics and Medicine	TIGEM	Naples Italy	Diego di Bernardo dibernardo@tigem.it
6	Istituto Italiano di Tecnologia	IIT	Milan Italy	Mattia Pelizzola Mattia.Pelizzola@iit.it
7	Association de la Recherche et le Dévéloppement des Méthodes et Processus Industriels	ARMINES	Paris France	Jean-Philippe Vert Jean-Philippe.Vert@mines.org
8	The University of Zurich	UZH	Zurich Switzerland	Mark Robinson mark.robinson@imls.uzh.ch
9	University of Cambridge	UCAM	Cambridge UK	Simon Tavaré simon.tavare@cruk.cam.ac.uk
10	University of Sheffield	USFD	Sheffield UK	Neil Lawrence n.lawrence@sheffield.ac.uk

4.2 Use and dissemination of foreground

Section A: Dissemination measures including scientific publications relating to foreground.

Our dissemination activities have remained in line with the plan from the original description of work (DoW) with no deviation (Section B3.2 of the DoW).

Training: We had a dedicated work-package for training which supported many successful and popular activities (listed in Table A2) including developer workshops, scientific meetings, training courses and scientific exchange visits. Teaching materials from these activities is available online. See report on WP9 for full details.

Dissemination of statistical software and scientific results: As well as the standard route of scientific publications (Table A1) and conference/seminar presentations (Table A2) we distributed a large number of open source statistical tools (Table B2) implementing our developed methods. We also had an entire work-package dedicated to improving the process of scientific software publication (Section 4.1.3, WP9) which will lead to better practice in software development, dissemination, usability and inter-operability. An important theme there was in the development of task-focused rather than method-focused documentation, which will make it easier for practitioners to adopt methods. Improve benchmarking (Section 4.1.3, WP10) is another area of importance so that users can select the optimal methods for a particular application.

Exploitation of project results and management of intellectual property: We have primarily been working on pre-competitive and open source developments, which is most fertile for rapidly developing and benchmarking new statistical tools. European biotechnology has benefitted from open collaboration (including our own SMEs) and a permissive choice of licensing which allows the inclusion of most R/Bioconductor tools within commercial products. For example, Genomatix has incorporated methods developed in this project in their licensed software Genomatix Mining Station. As well as the Bioinformatics sector represented by our SMEs, biotechnology companies developing new products (e.g. Oxford Nanopore) will greatly benefit from open source statistical tools to analyse their data, as has been the case in the past (e.g. with Affymetrix and Illumina).

Table A1 List of Scientific Publications

TEMPLATE A1: LIST OF SCIENTIFIC (PEER REVIEWED) PUBLICATIONS, STARTING WITH THE MOST IMPORTANT ONES										
NO.	Title	Main/First author	Title of the periodical or the series	Number, date or frequency	Publisher	Place of publication	Year of publication	Relevant pages	Permanent identifiers[16] (if available)	Is/Will open access[17] provided to this publication?
1	Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2	Love, M.	Genome Biology	Vol. 15, Issue, 12	BioMed Central	United Kingdom	2014	550	10.1186/s13059-014-0550-8	Yes
2	HTSeq--a Python framework to work with high-throughput sequencing data	Anders, S.	Bioinformatics	Vol 31, Issue 2	Oxford University Press	United Kingdom	2015	166-169	10.1093/bioinformatics/btu638	Yes
3	Robustly detecting differential expression in RNA sequencing data using observation weights	Zhou, X.	Nucleic Acids Research	Vol 42, Issue 11	Oxford University Press	United Kingdom	2014	2-10	10.1093/nar/gku310	Yes
4	Orchestrating high-throughput genomic analysis with Bioconductor	Huber, W	Nature Methods	Vol. 12, Issue 2	Nature Publishing Group	United Kingdom	2015	115-121	10.1038/nmeth.3252	Yes
5	Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays	Honkela, A.	Proceedings of the National Academy of Sciences of the United States	Vol. 113, Issue 42	National Academy of Sciences	United States	2015	13115-13120	10.1073/pnas.1420404112	Yes
6	Count-based differential expression analysis of RNA sequencing data using R and Bioconductor	Anders, S.	Nature Protocols	Vol. 8, Issue 9	Nature Publishing Group	United Kingdom	2013	1765-1786	10.1038/nprot.2013.099	Yes

[16] A permanent identifier should be a persistent link to the published version full text if open access or abstract if article is pay per view) or to the final manuscript accepted for publication (link to article in repository).

[17] Open Access is defined as free of charge access for anyone via Internet. Please answer "yes" if the open access to the publication is already established and also if the embargo period for open access is not yet over but you intend to establish open access afterwards.

7	BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach	Riebler, A.	Genome Biology	Vol. 15, Issue 2	BioMed Central	United Kingdom	2014	R35	10.1186/gb-2014-15-2-r35	Yes
8	HiC-Pro: an optimized and flexible pipeline for Hi-C data processing	Servant, N	Genome Biology	Vol. 16	Biomed Central	United Kingdom	2015	259-269	10.1186/s13059-015-0831-x	Yes
9	Efficient RNA isoform identification and quantification from RNA-Seq data with network flows	Bernard, E.	Bioinformatics	Vol. 30, Issue 17	Oxford University Press	United Kingdom	2014	2447-2445	10.1093/bioinformatics/btu317	Yes
10	Fast and accurate approximate inference of transcript expression from RNA-seq data	Hensman, J.	Bioinformatics	Vol. 31	Oxford University Press	United Kingdom	2015	3881-3889	10.1093/bioinformatics/btv483	Yes
11	INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments	de Pretis, S.	Bioinformatics	Vol 31, Issue 17	Oxford University Press	United Kingdom	2015	2829-2835	10.1093/bioinformatics/btv288	No
12	FourCSeq: analysis of 4C sequencing data	Klein, F. A.	Bioinformatics	Vol. 31. Issue 19	Oxford University Press	United Kingdom	2015	3085-3091	10.1093/bioinformatics/btv335	Yes
13	Inference of RNA Polymerase II Transcription Dynamics from Chromatin Immunoprecipitation Time Course Data	wa Maina, C.	PLoS Computational Biology	Vol. 10, Issue 5	Public Library of Science	United States	2014		10.1371/journal.pcbi.1003598	Yes
14	Enhancer loops appear stable during development and are associated with paused polymerase	Ghavi-Helm, Y.	Nature	Vol. 512, Issue 7512	Nature Publishing Group	United Kingdom	2014	96-100	10.1038/nature13417	Yes
15	Accounting for technical noise in single-cell RNA-seq experiments	Brennecke, P.	Nature Methods	Vol. 10, Issue 11	Nature Publishing Group	United Kingdom	2013	1093-1095	10.1038/nmeth.2645	Yes
16	Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis	Sabò, A.	Nature	Vol 511, Issue 7510	Nature Publishing Group	United Kingdom	2014	488-492	10.1038/nature13537	Yes

17	Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis	Cabezas-Wallscheid, N.	Cell Stem Cell	Vol. 15, Issue 4	Cell Press	United States	2014	507-522	10.1016/j.stem.2014.07.005	Yes
18	Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages	Ohnishi, Y.	Nature Cell Biology	Vol. 16, Issue 1	Nature Publishing Group	United Kingdom	2013	27-37	10.1038/ncb2881	Yes
19	Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression	Ay, F.	Genome Research	Vol. 24, Issue 6	Cold Spring Harbour Press	United States	2014	974-988	10.1101/gr.169417.113	Yes
20	Single-cell polyadenylation site mapping reveals 3' isoform choice variability	Velten, L	Molecular Systems Biology	Vol 11, Issue 6	Nature Publishing Group	United Kingdom	2015	812	10.15252/msb.20156198	Yes
21	Drift and conservation of differential exon usage across tissues in primate species	Reyes, A	Proceedings of the National Academy of Sciences of the United States	Vol. 110, Issue 38	National Academy of Sciences	United States	2013	15377-15382	10.1073/pnas.1307202110	Yes
22	Accurate identification of centromere locations in yeast genomes using Hi-C	Varoquaux, N.	Nucleic Acids Research	Vol. 43 Issue 11	Oxford University Press	United Kingdom	2015	5331-5339	10.1093/nar/gkv424	Yes
23	DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes	Melloni, G. E. M.	Genome Medicine	Vol. 6, Issue 2	BioMed Central		2014	1-13	10.1186/gm563	Yes
24	h5vc: scalable nucleotide tallies with HDF5	Pyl, P. T.	Bioinformatics	30	Oxford University Press	United Kingdom	2014	1-3	10.1093/bioinformatics/btu026	Yes
25	A statistical approach for inferring the 3D structure of the genome	Varoquaux, N.	Bioinformatics	Vol. 30, Issue 12	Oxford University Press	United Kingdom	2014	126-133	10.1093/bioinformatics/btu268	Yes
26	Multiple dimensions of epigenetic gene regulation in the malaria parasite	Ay, F.	BioEssays	Vol, 37, Issue 2	John Wiley and Sons Inc	United States	2015	182-194	10.1002/bies.201400145	Yes

27	Relationship between genome and epigenome - challenges and requirements for future research	Almouzni, G.	BMC Genomics	Vol. 15, Issue 1	BioMed Central	United Kingdom	2014	487	10.1186/1471-2164-15-487	Yes
28	Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C	Ay, F	BMC Genomics	Vol. 16	BioMed Central	United Kingdom	2015	121-137	10.1186/s12864-015-1236-7	Yes
29	Emerging bioinformatics approaches for analysis of NGS-derived coding and non-coding RNAs in neurodegenerative diseases	Guffanti, A.	Frontiers in Cellular Neuroscience	Vol. 8, Article 89	Frontiers Research Foundation		2014	1-10	10.3389/fncel.2014.00089	Yes
30	methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data	Kishore, K.	BMC Bioinformatics	Vol. 16, Issue 1	BioMed Central	United Kingdom	2015	313	10.1186/s12859-015-0742-6	Yes
31	A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples	Bernard, E	BMC Bioinformatics	Vol. 16	BioMed Central	United Kingdom	2015	262-271	10.1186/s12859-015-0695-9	Yes
32	Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters	Hensman, J	BMC Bioinformatics	No 14, Issue 1	BioMed Central	United Kingdom	2013	252	10.1186/1471-2105-14-252	Yes
33	Improved variational Bayes inference for transcript expression estimation	Papastamoulis, P.	Statistical Applications in Genetics and Molecular Biology		Berkelely Electronic Press	United States	2014	1-14	10.1515/sagmb-2013-0054	Yes
34	Computational and experimental methods to decipher the epigenetic code	de Pretis, S.	Frontiers in Genetics	Vol. 5	frontiersin.org	Switzerland	2014	1-6	10.3389/fgene.2014.00335	Yes
35	Computational epigenomics: challenges and opportunities	Robinson Mark	Frontiers in Genetics	Vol. 6	frontiersin.org	Switzerland	2015	88	http://dx.doi.org/10.3389/fgene.2015.00088	Yes
36	Selective transcriptional regulation by Myc: Experimental design and computational analysis of high-throughput sequencing data	Pelizzola, M.	Data in Brief	Vol. 3	Elsevier	United States	2015	40-46	10.1016/j.dib.2015.02.003	Yes

37	Transcriptome-wide Profiling and Posttranscriptional Analysis of Hematopoietic Stem/Progenitor Cell Differentiation toward Myeloid Commitment	Klimmeck, D.	Stem Cell Reports	Vol. 3, Issue 5	Cell Press	United States	2014	585-875	10.1016/j.stemcr.2014.08.012	Yes

Note regarding open access: All bar one of the publications listed is currently or will be open access (gold or green). Some publications are available via Europe PMC or PubMed Central.

Table A2 List of Dissemination Activities

TEMPLATE A2: LIST OF DISSEMINATION ACTIVITIES								
NO.	Type of activities[18]	Main leader	Title	Date	Place	Type of audience[19]	Size of audience (	Countries addressed
1	Public website	UNIMAN	RADIANT website	May 2013	http://radiant-project.eu	○ Scientific community;		International
2	Workshop	EMBL	SIG Frontiers in Somatic Variant Calling	Nov 2013	Heidelberg	○ Scientific community ○		International
3	Workshop	EMBL	Computational statistics for genome biology CSAMA 2013	June 2013	Brixen	○ Scientific community	60	International
4	Workshop	UCAM	European Bioconductor Developers' Meeting	Dec 2013	Cambridge	○ Scientific community	40	International
5	Open scientific meeting Workshop	UNIMAN	RADIANT General Meeting and Open Scientific Meeting	Jan 2014	Paris	○ Scientific community	40	International
6	Workshop	USFD	Genomic Workshop at MASAMB 2014	April 2014	Sheffield	○ Scientific community		International
7	Workshop/course	EMBL	Computational statistics for genome biology CSAMA 2014	June 2014	Brixen	○ Scientific community	60	International
8	Open scientific meeting Workshop	UNIMAN	RADIANT General Meeting and Open Scientific Meeting	July 2014	Heidelberg	○ Scientific community	30	International
9	Workshop	UNIMAN	Analysis of differential isoform usage by RNA-Seq: statistical methodologies and open software	Sept 2014	Strasbourg	○ Scientific community		International

[18] A drop down list allows choosing the dissemination activity: publications, conferences, workshops, web, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters, Other.

[19] A drop down list allows choosing the type of public: Scientific Community (higher education, Research), Industry, Civil Society, Policy makers, Medias ('multiple choices' is possible.

			workshop, ECCB 2014					
10	Workshop	UZH	Advanced R Software Carpentry Bootcamp	Nov 2014	Zurich	○ Scientific community	40	International
11	Workshop	EMBL	European Bioconductor Developers' Meeting	Jan 2015	Heidelberg	○ Scientific community	38	International
12	Open scientific meeting Workshop	UNIMAN	RADIANT General Meeting and Open Scientific Meeting	Jan 2015	Zurich	○ Scientific community	35	International
13	Workshop	UZH	Computational methods for high-dimensional single-cell mass cytometry data, BC2 conference	June 2015	Basel	○ Scientific community		International
14	Workshop/course	EMBL	Computational statistics for genome biology CSAMA 2015	June 2015	Brixen	○ Scientific community		International
15	Open scientific meeting Workshop	UNIMAN	RADIANT General Meeting and Open Scientific Meeting	July 2015	Naples	○ Scientific community		International
16	Workshop	UNIMAN	C1omics 2015: Single-cell omics methods and applications	Nov 2015	Manchester	○ Scientific community		International
17	Oral presentation	EMBL	Analysis and Integration of Transcriptome and Proteome Data	Nov 2015	C1omics Workshop, Manchester, UK	○ Scientific community	106	International
18	Oral presentation	EMBL	Shrinkage estimators in data analysis for comparative high-throughput sequencing experiments	April 2014	MASAMB 2015 Helsinki, Finland	○ Scientific community		International
19	Oral presentation	UNIMAN	Bayesian methods for transcript expression estimation and differential expression calling from RNA-Seq data	July 2015	Integrative RNA Biology SIG meeting, ISMB, Dublin, Ireland	○ Scientific community ○ Industry		International
20	Oral presentation	UNIMAN	Uncovering the dynamic response of breast cancer cells to estrogen with probabilistic models	July 2015	Next Generation Sequencing Data Congress, London, UK	○ Scientific community ○ Industry		International
21	Oral presentation	UNIMAN	Probabilistic modelling of omic	June 2015	Functional Genomics and Systems Biology	○ Scientific community ○ Industry		International

			time course data		Summer School, Hinxton, UK			
22	Oral presentation	UNIMAN	Gaussian process modelling for omic time course data	May 2015	Centre for Research in Statistical Methodology Seminars, Warwick	○ Scientific community		UK
23	Oral presentation	UNIMAN	Uncovering the mechanisms of transcription by integrating RNA-Seq and ChIP-Seq time course data	Nov 2014	Next Generation Sequencing (NGS, 2014)	○ Scientific community ○ Industry		International
24	Oral presentation	EMBL	Differential expression analysis for RNA-Seq using empirical Bayes priors for dispersion and fold change	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
25	Oral presentation	ARMINES	Flip-Flop: fast lasso-based isoform prediction from RNA-Seq data	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
26	Oral presentation	UNIMAN	Bayesian methods for transcript-level differential expression	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
27	Oral presentation	UZH	Robustly detecting differential expression in RNA sequencing data using observation weights	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
28	Oral presentation	IIT	Inference of mRNA production and degradation rates from time-course genomic data: is 29degradation actively re30gulated?	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
29	Oral presentation	TIGEM	Reverse-engineering the human retina gene regulatory networks from transcriptional inter-	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute,	○ Scientific community	45	International

			individual variability		Paris			
30	Oral presentation	EMBL	Comparing analysis tools for RNA-Seq: Thoughts on benchmarking approaches by simulation and with real data	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
31	Oral presentation	UZH	Simulation- and experimental-based strategies for benchmarking RNA-seq differential expression and splicing: A viewpoint	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
32	Oral presentation	UNIMAN	Overview of Arabadopsis data	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
33	Oral presentation	UCAM	How low can we go? - Reducing input quantities in RNA-seq	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
43	Oral presentation	EMBL	Statistical considerations for the analysis of single-cell RNA-Seq data	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
35	Oral presentation	ARMINES	Inferring the 3D structure of the genome from Hi-C data	Jan 2014	RADIANT General Meeting and Open Scientific Meeting, Marie Curie Institute, Paris	○ Scientific community	45	International
36	Oral presentation	EMBL	High resolution analysis of genome wide chromatin conformation data reveals stability of promoter enhancer interaction hubs during differentiation	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International

37	Oral presentation	UZH	ChIP-Seq without peak calling?	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International
38	Oral presentation	ARMINES	Isoform detection from multiple samples	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International
39	Oral presentation	TIGEM	Reverse-engineering the human retina transcriptome	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International
40	Oral presentation	UNIMAN	A tractable Gaussian Process model to identify the perturbation point and score differential expression from time course data	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International
41	Oral presentation	ARMINES	Inferring the 3D structure of the genome from Hi-C data	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	30	International
42	Oral presentation	GENOMNIA	Genome Browser Software demonstration	July 2014	RADIANT Project Meeting and Open Scientific Meeting, EMBL, Heidelberg	○ Scientific community	25	International
43	Oral presentation	EMBL	Single cell RNA-Seq analysis of medullary thymic epithelial cells (mTECs) reveals patterns of promiscuous gene expression in the thymus	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
44	Oral presentation	GENOMATIX	A reliability check for NGS Mapping Results	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
45	Oral presentation	ARMINES	Isoform inference from multiple RNA-seq samples	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting,	○ Scientific community	30	International

					University of Zurich			
46	Oral presentation	TIGEM	Reverse-engineering the human retina transcriptome	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
47	Oral presentation	ARMINES	Centromere identification from Hi-C data	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
48	Oral presentation	GENOMATIX	Transcript annotations – same same but different	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
49	Oral presentation	UZH	CRISPR/Cas9 variant analysis	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
50	Oral presentation	IIT	INSPECT: Infer MRNA Sythesis Process and Degradation Rates and Dynamics from RNA – and 4sU – seq time course	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
51	Oral presentation	UNIMAN	Analysis of splicing-associated delays in transcription	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
52	Oral presentation	USFD	Spike and Slab GPLVM for Extracting Regulator Activity Profiles	Jan 2015	RADIANT Project Meeting and Open Scientific Meeting, University of Zurich	○ Scientific community	30	International
53	Oral presentation	IIT	The regulation of RNA dynamics following Myc overexpression	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
54	Oral presentation	UNIMAN	Integrating ChIP-seq time course data to infer regulatory enhancer-promoter interactions	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International

55	Oral presentation	GENOMNIA	The dynamic visualization of Hi_Seq dataset	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
56	Oral presentation	UZH	Detecting differentially spliced genes - a comparison of counting methods and an open benchmarking platform for the community	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
57	Oral presentation	EMBL	Single-cell RNA-Seq analysis to unravel the patterns and mechanisms of promiscuous gene expression in thymic epithelial cells	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
58	Oral presentation	ARMINES	How cells know where they are: building a transcriptomic map of the frog embryonic ectoderm	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
59	Oral presentation	USFD	An unsupervised approach for tumor phenotyping from tissue microarray images of prostate cancer	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
60	Oral presentation	UCAM	Quality assessment tools for nanopore sequencing	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
61	Oral presentation	TIGEM	Transcriptome analysis of human retina	Jul 2015	RADIANT General Meeting and Open Scientific Meeting, TIGEM, Naples	○ Scientific community	30	International
62	Poster	Genomnia	Radiant Genome Browser – exploring methylseq datasets in a genome browser	Oct. 2015	15th International Workshop on Network Tools and Applications for Biology (Nettab), Bari, Italy	○ Scientific community		International

63	Poster	Genomnia	Radiant Genome Browser – exploring methylseq datasets in a genome browser	June 2015	12th Annual Meeting of the Bioinformatics Italian Society, Milan, Italy	○ Scientific community		International
64	Poster	EMBL	Unravelling patterns of ectopic gene expression in medullary thymic epithelial cells	Sep 2015	Single-cell Genomics Conference 2015 Utrecht, Netherlands	○ Scientific community		International
65	Website	UZH	Differential splicing comparison code		https://github.com/markrobinsonuzh/diff_splice_paper	○ Scientific community		International
66	Website	UCAM	Genomic Variation Detection		https://sites.google.com/site/benchmarkdatasets/	○ Scientific community		International
67	Website	UZH	DNA Methods Methylation Analysis		http://radiant-project.eu/Output/Benchmarking/Methylation.html	○ Scientific community		International
68	Website	UZH	Simulated RNA-Seq data		https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3766/	○ Scientific community		International
69	Website	UNIMAN	Transcript abudndance code		https://github.com/BitSeq/BitSeqVB_benchmarking/wiki	○ Scientific community		International
70	Flyer	UNIMAN	RADIANT Flyer	May 2014	Downloadable from RADIANT website	○ Scientific community		International

Section B: Exploitable foreground and plans for exploitation
This information will be made public unless otherwise stated

Part B1 Application for patents, trademarks etc if applicable (in table form)

No applications were made for patents or trademarks.

Part B2

Purpose of RADIANT foreground: to advance knowledge and provide access to enhanced methodologies.
All RADIANT foreground is or will be available for exploitation to the scientific community.

Type of Exploitable Foreground[20]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[21]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
General advancement of knowledge	An R package providing standard formatting styles for vignettes and other Bioconductor documents	No	N/A	BiocStyle software	J58.2 Software publishing	Already available	None	No 'owner'. Involvement of EMBL
General advancement of knowledge	An R package targeted for transcript expression analysis and differential expression analysis of RNA-seq data in a two stage process.	No	N/A	BitSeq ssoftware	J62.0.1 Computer programming activities	Already available	None	No 'owner'. Involvement of UNIMAN
General advancement of knowledge	R tools for computational epigenomics developed for the analysis, integration and simultaneous visualization	No	N/A	CompEpiTools software	M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvement of

[1] A drop down list allows choosing the type of foreground: General advancement of knowledge, Commercial exploitation of R&D results, Exploitation of R&D results via standards, exploitation of results through EU policies, exploitation of results through (social) innovation.
[21] Select (NACE nomenclature) : http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

Type of Exploitable Foreground[20]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[21]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
	of various (epi)genomics data types across multiple genomic regions in multiple samples.							
General advancement of knowledge	An R package for inference of differential expression from count data from high-throughput sequencing assays.	No	N/A	DESeq 2 software	J58.2 Software publishing J62.0.1 Computer programming activities M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvement of EMBL
General advancement of knowledge	A general purpose R package for detection of differential features starting from a count table (e.g., RNA-seq, ChIP-seq, SAGE and CAGE, metagenome counts, transcript-level estimated counts).	No	N/A	edgeR Software	J58.2 Software publishing	Already available	None	No 'owner'. Involvement of UZH
General advancement of knowledge	A general purpose R package for detection of differential features starting from a count table (e.g., RNA-seq, ChIP-seq, SAGE and CAGE, metagenome counts, transcript-level estimated counts).	No	N/A	edgeR software	J62.0.1 Computer programming activities	Already available	None	No 'owner'. Involvement of UZH
General	An R package aimed to	No	N/A	flipflop	M72.1.1 Research	Already available	None	No 'owner'.

Type of Exploitable Foreground[20]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[21]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
advancement of knowledge	discover which isoforms of a gene are expressed in a given sample together with their abundances, based on RNA-Seq read data.			software	and experimental development on biotechnology			Involvement of ARMINES
General advancement of knowledge	An R package dedicated to the analysis of 4C sequencing data, possibly multiplexed	No	N/A	FourCSeq software	J58.2 Software publishing	Already available	None	No 'owner'. Involvement of EMBL
General advancement of knowledge	Gaussian processes framework for Python.	No	N/A	GPy software	J62.0.1 Computer programming activities	Already available	None	No 'owner'. Involvement of USFD
General advancement of knowledge	R package containing experimental data and complete executable transcript (of statistical analysis presented in "Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages" by Y. Ohnishi, et al Nature Cell Biology (2014) 16(1): 27-37	No	N/A	Hiiragi2013 software	M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvement of EMBL
General advancement of knowledge	An R package for the integrative analysis of RNA- and 4sU-seq data to study the dynamics of transcriptional regulation	No	N/A	INSPEcT software	J58.2 Software publishing	Already available	None	No 'owner'. Involvement of IIT

Type of Exploitable Foreground[20]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[21]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
General advancement of knowledge	Tools for the quality assessment of Oxford Nanopore MinION data in R	No	N/A	IONiseR software	J58.2 Software publishing	Already available	None	No 'owner'. Involvement of UCAM
General advancement of knowledge	R package for memory efficient analysis of base resolution DNA methylation data in both the CpG and non-CpG sequence context.	No	N/A	methylPipe software	J62.0.1 Computer programming activities	Already available	None	No 'owner'. Involvement of IIT
General advancement of knowledge	Python code used to study the dynamicsc of RNA polymerase II transcription from chromatin immunoprecipitation time course data.	No	N/A	PyPol-II software	M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvement of Genomatix, USFD and UNIMAN
General advancement of knowledge	Visualization tool for methyl seq analysed data from the MethylPipe Package	No	N/A	RADIANT Genome Browser	J58.2 Software publishing	Already available	None	No 'owner'. Developed by Genomnia
General advancement of knowledge	R tools for the analysis of enrichment-based epigenomic data.	No	N/A	Repitools software	J62.0.1 Computer programming activities	Already available	None	No 'owner'. Involvement of UZH
General advancement of knowledge	R package for two-sample time series analysis using Gaussian process methods	No	N/A	DETime software	M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvment of UNIMAN

Type of Exploitable Foreground[20]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[21]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
General advancement of knowledge	Optimized and flexible pipeline for processing Hi-C data from raw reads to normalized contact maps	No	N/A	Hic-Pro software	J58.2 Software publishing	Already available	None	No 'owner'. Involvment of ARMINES
General advancement of knowledge	Python software to jointly infer the locations of all centromeres in a single yeast genome from Hi-C data	No	N/A	Centurion software	J58.2 Software publishing J62.0.1 Computer programming activities M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvment of ARMINES
General advancement of knowledge	Python software to infer the 3D structure of a genome from Hi-C data	No	N/A	Pastis	J58.2 Software publishing J62.0.1 Computer programming activities M72.1.1 Research and experimental development on biotechnology	Already available	None	No 'owner'. Involvment of ARMINES

4.3 Report on societal implications

Please refer to online submission