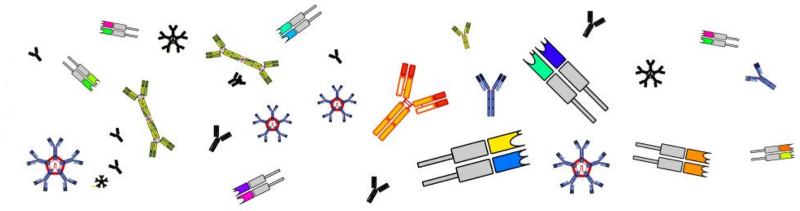


DELIVERABLE 4.1

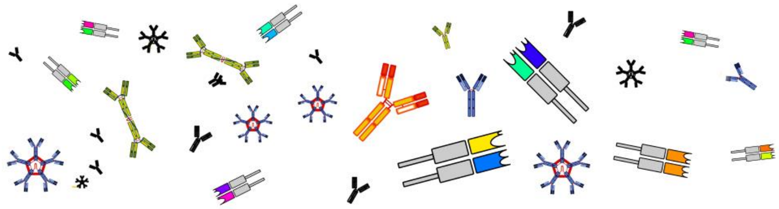
CRITICAL ANALYSIS CAPABILITIES AND TOOLS IDENTIFIED FOR INTEGRATION

WORK PACKAGE NUMBER: WP4

WORK PACKAGE TITLE: ANALYSIS PIPELINES

TYPE: REPORT

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



Document Information

iReceptor Plus Project Information	
Project full title	Architecture and Tools for the Query of Antibody and T-cell Receptor Sequencing Data Repositories for Enabling Improved Personalized Medicine and Immunotherapy
Project acronym	iReceptor Plus
Grant agreement number	825821
Project coordinator	Prof. Gur Yaari
Project start date and duration	1st January, 2019, 48 months
Project website	http://www.ireceptor-plus.com

Deliverable Information	
Work package number	WP4
Work package title	Analysis Pipelines
Deliverable number	D4.1
Deliverable title	Critical Analysis Capabilities and Tools Identified for Integration
Description	Assessment of analysis capabilities, analysis tools, visualization capabilities and technological considerations to drive the development of platform-tool integration mechanisms for advanced analysis of AIRR-seq data.
Lead beneficiary	University of Texas Southwestern Medical Center
Lead Author(s)	Scott Christley

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Contributor(s)	All beneficiaries
Revision number	5
Revision Date	June 21, 2019
Status (Final (F), Draft (D), Revised Draft (RV))	RV
Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	PU

Document History

Revision	**Date**	**Modification**	**Author**
1	May 29, 2019	Initial version	Scott Christley
2	June 3, 2019	Review	Brian Corrie
2	June 10, 2019	Review	Bracha Ehrman
2	June 11, 2019	Review	Felix Breden
2	June 14, 2019	Incorporate review comments, write section 2	Scott Christley
3	June 20, 2019	Review	Uri Hershberg
4	June 20, 2019	Correct name	Milena Mirkis
5	June 21, 2019	Review	Sarah Taylor

Approvals

	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Prof. Gur Yaari	Bar Ilan University	June 30, 2019	GY
WP Leaders	Dr. Lindsay Cowell	University of Texas Southwestern Medical Center	June 30, 2019	LC

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Table of Contents

Executive Summary 5

Deliverable description 6

1. Analysis Capabilities 7

- 1.1. Basic Analysis 7
- 1.2. CDR3 Analysis 8
- 1.3. Clonal Analysis 8
- 1.4. Diversity Metrics 9
- 1.5. Repertoire Similarity / Distance / Overlap 9
- 1.7. Mutational Analysis 10
- 1.8. Lineage Trees 11
- 1.9. Machine Learning Features (WP5 integration) 11
- 1.10. Systems Immunology (WP6 integration) 12
- 1.11. Single-Cell Analysis (WP7 integration) 12

2. Tool Capability Matrix 12

3. Visualization 14

4. Technology Considerations 14

Summary and future work 15

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Executive Summary

One of the goals of the iReceptor Plus project is to make AIRR-seq data Findable, Accessible, Interoperable and Reusable, or FAIR. The iReceptor Plus project makes data findable and accessible by providing a platform that can search and federate data from multiple studies across multiple repositories around the world. One of the key challenges is making the data reusable. For data to be reusable, it must be possible to redo analyses (for scientific reproducibility) as well as perform new analyses. WP4 brings together the expertise from leading groups in the development of analysis pipelines to design and implement an advanced analysis platform in which it is simple to integrate and apply existing and novel AIRR-seq data analysis tools to complex, federated AIRR-seq data.

Key tasks in WP4 include a detailed assessment of the AIRR-seq data analysis tool landscape (Task 4.1), the development of platform-tool integration mechanisms for advanced analysis and integration of analysis tools (Task 4.2 and 4.5), the standardization of analysis output formats for tool interoperability adn to be in line with community standards (Task 4.3), and improvements to the iReceptor Plus Graphical User Interface for analysis tool selection, running analysis tools, analysis job management and monitoring, and presenting and visualizing analysis results (Task 4.4).

This deliverable D4.1 signifies the successful completion of Task 4.1 with a detailed assessment of the AIRR-seq data analysis tool landscape. This deliverable incorporated feedback from all beneficiaries across multiple meetings and discussions. Furthermore, analysis requirements from WP5, WP6 and WP7 were incorporated to insure a consistent and integrated approach across the whole iReceptor Plus project. The results of the assessment identified over fifty analysis capabilities divided among eleven categories. Numerous existing analysis tools were identified but notably a number of analysis capabilities are not performed by existing tools. This implies that new development is required to provide those capabilities to users of iReceptor Plus. Furthermore, technological considerations have identified that some tools may have scalability issues to handle large-scale data and/or the distributed and federated architecture of the AIRR repositories. Lastly, visualization capabilities were assessed in anticipation of improvements to the iReceptor Plus GUI as part of Task 4.4.

Moving forward, the results of this assessment need to be consolidated with the use case scenarios identified as part of WP1. This will provide overall requirements that will drive the design and implementation of iReceptor Plus. Lastly, a subset of analysis capabilities will be identified to be prototyped in the development of the platform-tool integration mechanisms as part of Task 4.2.

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Deliverable description

Deliverable D4.1 is the main result of the work done in WP4 Task 4.1.

The objective of this task is to characterize the analysis capabilities that iReceptor Plus will provide to users, and then perform an assessment of existing analysis tools that can provide those capabilities. As part of that assessment, technological considerations were examined in the context of the federated data repository architecture and large-scale data. Also, visualization capabilities for presenting analysis results to users were considered in anticipation of improvements to the iReceptor Plus GUI.

In the following sections, we describe the set of analysis capabilities that iReceptor Plus will provide. The capabilities are split among eleven categories (sections 1.1 - 1.11) including three categories that represent analysis integration with work packages WP5, WP6 and WP7. For each category, we provide a basic description of the analyses as well as some biological information that supplements the motivation for those analyses. However, we don't attempt to describe each analysis method in detail. Many analyses can be used for both B and T cell repertoires. However, B cells undergo additional biological processes of somatic hypermutation and affinity maturation, which entails additional analysis methods specific to them. In the category description, we mention if the analyses are specific to B cells.

In several cases we have noted issues or challenges regarding the analysis category. The old saying, "garbage in, garbage out", is definitely applicable to AIRR-seq data. The federated architecture of multiple AIRR-seq studies spread out across multiple data repositories introduces concerns regarding all aspects of data quality: validity, accuracy, completeness, consistency and uniformity. In particular, studies will utilize different experimental protocols, sequencing technologies and data processing protocols. Naive comparison of AIRR-seq studies with differing protocols and technologies can uncover variations (or similarities) that do not represent true biological variation (or similarity) but are actually artifacts of comparing data sets based on different protocols. The iReceptor Plus project is cognizant of these challenges and will develop mechanisms to ameliorate them so that users are performing analyses with the highest quality data possible.

Lastly, the computational architecture for calculating these analyses is still to be defined. The characterization of the analysis capabilities has provided a set of requirements that will guide the design and implementation of that computational architecture. Task 4.2 of WP4 will develop the platform-tool integration mechanisms for advanced analysis, and in the process, various computational architectures will be explored, prototyped and evaluated.

1. Analysis Capabilities

1.1. Basic Analysis

Basic analysis considers gene usage and distribution for the V, D and J genes at the level of alleles, genes, gene families, and locus types. Usage and distributions are considered for singular genes, gene combinations, productive and non-productive sequences, within individuals, between groups, and across populations. The list of basic analysis capabilities that iReceptor Plus will provide includes:

- V gene usage
- J gene usage
- D gene usage
- V-J gene joint distribution
- Allele frequency
- Isotype distribution
- Fraction of productive sequences
- Genotype
- Haplotype

Gene usage analysis is reliant upon a complete and well-curated germline gene set. For example, analysis of B cells with incorrect allele assignment can introduce false mutations that bias downstream analyses. However, it is clear from recent studies that the current germline gene databases are incomplete. In response, the AIRR Community has formed the Germline Database Working Group (GLDB) to consider these issues and has formed the Inferred Allele Review Committee (IARC) to review and accept novel alleles inferred from AIRR-seq data. Key members of the iReceptor Plus project are associated with the GLDB, attend the periodic meetings, and will keep iReceptor Plus informed of its activities. The GLDB is working closely with the International Immunogenetics Information System (IMGT), which is the currently accepted database of germline genes.

As part of iReceptor Plus, we will adopt a standard germline gene set from the IMGT database to ensure that analyses are comparable across studies and repositories. Periodically, the germline gene set will be updated to incorporate new information from the IMGT database. Versioning will be utilized to prevent incompatible comparisons as well as to maintain scientific reproducibility.

It is important to note that even with accurate germline gene sets the identification of specific distinct germlines gene usage can be difficult because (1) Some gene segments (V, D or J) are very similar; (2) some experiments have only partial reads that do not cover the variable CDR1 and CDR2 regions and (3) some B cell receptor sequences are highly mutated. This can lead to V, D or

J ties where two similar gene segments cannot be distinguished at some level of confidence. This problem is doubly true at the allele level which cannot be distinguished for many genes. In those particular cases, some of the analyses mentioned above, e.g. allele frequency, cannot be accurately provided, and iReceptor Plus will notify users when those situations arise.

1.2. CDR3 Analysis

The complementary determining region 3 (CDR3) of a receptor chain is a highly variable sequence segment. Its variable nature leads to the high diversity of immune repertoires and is one of the key determinants in the recognition of antigens by T cells and B cells. The list of CDR3 analysis capabilities that iReceptor Plus will provide includes:

- CDR3 length distribution
- CDR3 nucleotide and amino acid composition
- Biochemical properties
- CDR3 sharing/uniqueness between repertoires
- Antigen-specific T cell and B cell receptors.

1.3. Clonal Analysis

When a T cell or B cell divides to form two daughter cells, both daughter cells will have identical receptor chain sequences and thus identical antigen specificities. Further subdivisions will create additional identical daughter cells. This is a hallmark of the adaptive immune system as it allows for an exponential increase in the number of cells that mounts a highly specific immune response to a particular pathogen. The resulting population of cells, all arising from the same founding parent cell, is a clone. For B cells, somatic hypermutation in the germinal center alters the receptor chain sequences of daughter cells with the intent of increasing the affinity for the antigen. Even though their receptor chain sequence is different, these higher-affinity B cells are still considered part of the same clone as they are descendants of the same parent cell. The parent B cell and its descendants form a clonal lineage which is described further in Section 1.8 Lineage Trees.

One of the primary tasks of immune repertoire analysis is determination of clones from the sequencing data, and clones are the basis for many other analyses. However, clones are not directly observable from AIRR-seq data. Instead a key element in the comparison of the behaviour of immune cell populations is the definition of individual clonotypes. A clonotype is an operational definition or inference of a clone from the sequencing data. Example definitions are (1) same V gene, same J gene, same CDR3 length, 85% homology between sequence, or (2) same V gene, same J gene, same CDR3 amino acid sequence. The definitions can vary based upon the cell type, experimental conditions, error rates of sequencing platforms, disease characteristics, or other considerations by the researcher. For this reason, iReceptor Plus will provide the capability for flexible clonotype definitions. The list of clone analysis capabilities that iReceptor Plus will provide includes:

- Flexible clonotype definitions
- Clone ranked abundance distribution
- Clone frequencies
- Clone size
- Clonal shift and clonal drift
- Public clone analysis across people

Determination of clones and their abundance from immune repertoire sequencing data can be challenging. Specifically due to the nature of clones having identical receptor sequences, it can be hard to distinguish whether sequences are duplicated due to the experimental and sequencing protocol versus whether they come from separate cells. Some experimental protocols incorporate mechanisms that allow for quantitative assessment of clonal abundance, while other experimental protocols do not, leading them to be semi-quantitative. Furthermore, sequencing error can cause sequences to be considered separate clones when they should be included in existing clones. The iReceptor Plus project is cognisant of these issues and will utilize indicators and warnings to inform users when particular analyses might be affected.

1.4. Diversity Metrics

Diversity metrics are quantitative measures that reflect how many different types are present in a data set. For AIRR-seq data, the clone is the type often considered but also other categories such as genes, gene families, and functional, genotypic and phenotypic characteristics may be utilized. The data set may be a single repertoire or multiple repertoires that encompass groups or populations. The list of diversity metric capabilities that iReceptor Plus will provide includes:

- Diversity at different Renyi/Hill numbers from zero (richness) and up to infinity (Berger-Parker)
- Rarefaction curves and sampling sufficiency estimates

1.5. Repertoire Similarity / Distance / Overlap

An individual's immune repertoire is constantly changing as the adaptive immune system responds to new threats as well as diminishes the response to successfully cleared challenges. Various measures have been utilized to assess the similarity, distance and overlap of one repertoire with another. Comparative analysis can be between repertoires from the same individual, repertoires from different individuals, and repertoires from different groups or populations. The list of repertoire similarity, distance and overlap measures that iReceptor Plus will provide includes:

- Morisita Horn
- Bray-Curtis
- Jaccard
- Venn diagrams
- Cosine similarity
- Levenstein or Hamming distance to build network
- Bipartite graphs
- Differential clone expression

1.7. Mutational Analysis

When B cells undergo somatic hypermutation, mutations can occur throughout the V, D and J gene segments. T cells do not undergo somatic hypermutation so mutational analyses are not typically performed on them. Determination of mutations is performed by comparing the receptor sequence with the germline gene sequence. Mutations can be synonymous whereby the altered nucleotide does not cause a substitution of the amino acid in the corresponding translated receptor protein, and presumably does not alter the receptor's function; while non-synonymous mutations do cause a substitution of the amino acid and potentially changes how the receptor binds and the cell functions. Affinity maturation exerts selection pressure on functional receptors to select mutations that increase the affinity of the B cell receptor for its antigen. Mutations of non-functional receptors are assumed to be neutral, meaning they do not change the fitness of the receptor, and the comparison of mutations between functional and non-functional receptors can be performed to better understand the relationship between selection pressure and engineered mutation patterns of the V gene. However, it is often not directly determinable whether a particular rearrangement sequence is functional without performing additional biological experiments. Instead, an operational definition of productive (or non-productive) is utilized for bioinformatic analysis. The IMGT definition for a productive rearrangement is (1) the coding region has an open reading frame, (2) no defect in the start codon, splicing sites or regulatory elements, (3) no internal stop codons, and (4) an in-frame junction region; otherwise the rearrangement is considered non-productive. The list of mutational and selection analysis capabilities that iReceptor Plus will provide includes:

- Mutation frequency
- Substitution frequency
- Selection N/S index
- Baseline for quantification of selection
- Clone diversity analysis by position AA usage

1.8. Lineage Trees

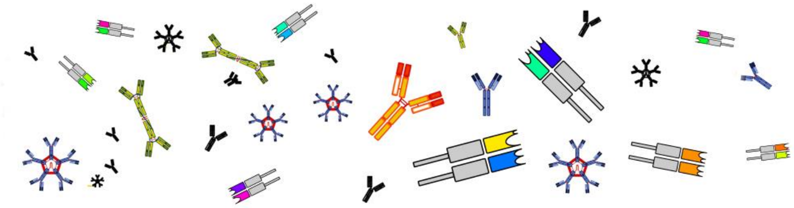
Lineage trees are another set of B cell specific analyses. A B cell clonal lineage is the ancestral B cell and all of its offspring. The analysis of lineage trees can be particularly important for the study of B cell clones, because when B cells undergo somatic hypermutation, their receptor sequence changes, and these mutations can be used to reconstruct the evolutionary history of the clonal lineage (i.e., lineage tree). The root of the tree corresponds to the original ancestral B cell, internal nodes represent intermediate B cells, and the leaves represent the current high-affinity B cells. The shape and structure of the lineage tree can be analysed to better understand the temporal progression of the clonal lineage as it evolves higher affinity for the antigen (often a viral or bacterial pathogen). Multiple algorithms exist to calculate the lineage tree with various parameters that require adjusting to optimize the results. The list of lineage tree analysis capabilities that iReceptor Plus will provide includes:

- Inference of lineage tree utilizing multiple algorithms with adjustable parameters
- Association of lineages with gene co-expression analysis of specific genes
- Association of metadata on lineages
- Mutation and selection analysis on the lineage

1.9. Machine Learning Features (WP5 integration)

WP5 of iReceptor Plus has the objective to develop, adapt and implement advanced algorithms for mining AIRR-seq data. These advanced algorithms utilize features of the data which are individual measurable properties or characteristics of the data or the phenomenon represented by the data. While features can be existing data fields and variables, many are derived values intended to be informative and non-redundant. An objective of WP4 is to calculate these features either as needed or with pre-computation and caching on data queried from AIRR repositories by utilizing the analysis pipelines as part of the overall computational architecture of iReceptor Plus. The list of machine learning features that iReceptor Plus will provide includes:

- k-mers/n-mers
- Receptor abundance
- Mutability model
- Lineage characteristics, distributions and shapes
- Subject attributes
- Sample attributes
- Cell attributes
- Hierarchical clustering
- Motifs



1.10. Systems Immunology (WP6 integration)

WP6 of iReceptor Plus has the objective to provide integrative analysis of AIRR-seq data with other "omics" and clinical data. An objective of WP4 is to provide the overall computational architecture to facilitate these integrative analyses. The list of systems immunology analysis capabilities that iReceptor Plus will provide includes:

- Immune cell phenotyping
- Subject attributes
- Gene expression
- Serum cytokines expression
- Auto-antibodies
- Microbiota composition and abundance

1.11. Single-Cell Analysis (WP7 integration)

WP7 of iReceptor Plus has the objective to host single-cell AIRR-seq data and to support queries on that data including phenotypic and functional attributes. Single-cell sequencing represents a substantial advance compared to current bulk sequencing of B and T cells. Specifically, it allows for both chains (e.g., heavy and light for B cells, and α and β for T cells) to be sequenced, it can be combined with transcriptomics or with flow cytometry data describing the cell phenotype, and it can be combined with experimental measurements of a cell's individual receptor reactivity. Many of the preceding sections describing analysis capabilities are in the context of bulk sequencing of a single receptor chain. Thus, an objective of WP4 is to also provide those analytical capabilities in the context of paired-chain single cell sequencing. The list of single cell analysis capabilities that iReceptor Plus will provide includes:

- Chain pairing: provide many of the analyses for bulk sequencing of single chains at the level of coupled chains.
- Clustering and subset based on gene expression patterns
- Clustering and subset based on other phenotype indicators
- Clustering and subset based on receptor reactivity
- Tree building under the assumption of drop out
- Association of lineages with gene co-expression analysis of specific genes

2. Analysis Tool Assessment

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

The previous sections have characterized the analysis capabilities that iReceptor Plus will provide to users, and in this section, we perform an assessment of existing analysis tools that can provide those capabilities. Further assessment regarding technical considerations for tools is addressed in Section 4. We consider tools internal to the iReceptor Plus consortium members, and thus have direct control over their design, development and implementation, as well as tools external to the iReceptor Plus consortium. For external tools, we primarily consider tools available under an open source license which allows us to modify their source code, if necessary, to integrate with iReceptor Plus. We did not attempt to assess all existing analysis tools, as many have duplicate capabilities, but instead focused on tools with broad adoption and are well supported. Internal tools considered include:

- ImmuneDB
- VDJServer
- VDJPipe
- RepCalc
- sciReptor
- iReceptor

External tools considered include:

- IgBlast
- pRESTO
- ChangeO
- Allakazam
- Shazam
- TIgGER
- RAbHIT

Of the external tools, IgBlast, pRESTO, ChangeO, Allakazam and Shazam have already been integrated into VDJServer, which will allow them to be easily integrated into iReceptor Plus.

With the set of tools considered, both internal and external, we have broad coverage of all eleven analysis categories, and within each category at least 80% of the analysis capabilities are provided. The exceptions are the three categories, Machine Learning Features, Single-Cell Analysis and Systems Immunology, associated with WP5, WP6 and WP7, which is to be expected as these are novel capabilities being provided by iReceptor Plus. For the approximately 20% of the analysis capabilities not covered, we will either enhance existing tools or consider integrating additional external tools that provide that capability. Overall, this is a positive assessment as it implies that significant analysis capabilities will be available to iReceptor Plus users as these tools are integrated. Task 4.2 of WP4 will develop the platform-tool integrations mechanisms for advanced analysis, and these tools will be integrated as part of that task. Task 4.5 of WP4 will perform a second phase of tool assessment to determine additional tools to be integrated into iReceptor Plus.

3. Visualization

Visualization is a vital aspect for users to understand and explore their analysis results. A portion of Task 4.4 for WP4 involves the development of a graphical user interface for iReceptor Plus for presenting and visualizing analysis results. Our objective is to provide flexible and customizable visualization capabilities while providing integration mechanisms with other platforms and tools when more sophisticated techniques or publication quality is needed. The list of visualization capabilities that iReceptor Plus will provide includes:

- Heatmap
- Histograms
- Line plots
- Dot plots
- Bar plots
- Chord graphs
- Tree visualization
- Circos graphs
- Logo graphs
- Network/Graph plots
- PCA/MDS/t-SNE plots
- Tableau integration
- Cytoscape integration

4. Technology Considerations

Scalability is one of the primary concerns when considering what tools, technologies and algorithms are to be implemented in iReceptor Plus. The federated architecture of multiple data repositories implies horizontal scaling for AIRR-seq data that already includes billions of data records. Based upon current estimates, this will increase by billions more every year with the potential to approach peta-scale levels over the coming decades. Horizontal scaling means that more nodes (computers, repositories) are added to the system. This is in contrast to vertical scaling where more resources (compute, memory) are added to a single node. The distributed architecture of the data repositories suggests that iReceptor Plus should develop a computational analysis architecture that is also distributed and can horizontally scale. Unfortunately, not all analysis tools have been implemented to scale horizontally. Furthermore, while some tools might be re-implemented with different algorithms, not all analysis methods have algorithms that can be scaled horizontally. This implies that the iReceptor Plus project will need to also provide vertical scaling in order to provide some analysis capabilities to users.

In assessing the set of analysis tools, besides characterizing what analysis capabilities the tools provide, we also characterize the technologies utilized by the tools. These technology considerations will allow the iReceptor Plus project to evaluate whether particular tools can be used as is, can be modified for scalability, or can not be utilized. The list of technological considerations includes:

- Programming language
- License: open source or proprietary
- Parallelization
    - multi-threaded
    - multi-processor
    - MPI or OpenMP
    - map-reduce
- Stream or load input data
- Support for AIRR Community standards
    - AIRR TSV format for rearrangement data
    - AIRR metadata format for repertoire data
    - AIRR germline set format
    - AIRR software working group guidelines
- Docker image, continuous integration, automated builds
- Command-line interface for batch processing
- Graphical user interface
- Generate visualizations (charts, figures, graphs, plots)
- Test suite
- Requires or utilizes a database
- Provides REST API

Summary and future work

This deliverable D4.1 signifies the successful completion of Task 4.1 with a detailed assessment of the AIRR-seq data analysis tool landscape. This deliverable incorporated feedback from all beneficiaries across multiple meetings and discussions. Furthermore, analysis requirements from WP5, WP6 and WP7 were incorporated to insure a consistent and integrated approach across the whole iReceptor Plus project. The results of the assessment identified over fifty analysis capabilities divided among eleven categories. The combination of analysis capabilities, analysis tools, visualization capabilities and technological considerations will drive the design and development of iReceptor Plus.

The computational architecture of iReceptor Plus to calculate and provide these analysis capabilities to users is still to be defined. Task 4.2 of WP4 will develop the platform-tool integration mechanisms for advanced analysis, and in the process, various computational architectures will be explored, prototyped and evaluated. Various subtasks will be performed as part of the overall Task 4.2. We have already identified a number of these subtasks and describe them here as future work to be performed:

1. Each analysis method, and more specifically the algorithm, needs to be categorized in regard to its scalability. Can it be parallelized and using which parallel techniques?
2. The list of analysis capabilities needs to be consolidated with the use case scenarios generated by WP1. The use cases describe the end-to-end usage of iReceptor Plus that includes more than just analysis. This consolidation will help drive development of analysis workflows and the graphical user interface.
3. Select a limited set of analysis capabilities and analysis tools to prototype in various computational architectures. This will allow each architecture to be evaluated in terms of performance and scalability and to expose potential bottlenecks. The selected set of analysis capabilities should include a broad spectrum of algorithmic types so that scalability concerns can be properly evaluated. Furthermore, data sizes ranging from very small to very large will need to be tested and evaluated.
4. The analysis output for each capability needs to be assessed in terms of its size and structure. Is the output too large to easily fit into memory of a web browser for visualization? Can the output be split such that partial results can be shown? Can the output size vary greatly thus requiring logic to dynamically check? Is the structure sufficiently complex that parsing functions are needed to interpret the data?