ACGT Context Scenario (Biostatisticians)

Context of Use	Dialogue Principle	System Requirements
Introduction Thierry and Francesca are scientists working in the field of bioinformatics and biostatistics. They are involved in the European project ACGT – Advancing Clinico-Genomic Clinical Trials on Cancer. The aim of this project is to develop a powerful European Biomedical Grid infrastructure on cancer for sharing data and data-processing methods and tools, and especially to design and develop interoperable and smoothly integrated data mining software components and tools for clinicians, biomedical researchers and statisticians. Among the biostatisticians' tasks, the ones that are relevant for ACGT are: - contribute to the design of biological experiments and clinical studies/trials involving genomic data - suggest the appropriate statistical methods for the analysis when existing methods are available - assess algorithms and methodologies for the above-described analyses - contribute to the development/extension of algorithms and methodologies - design workflows for the analyses from the biological	Suitability for the task Controllability Self-descriptiveness Suitability for the task Suitability for the task	A workflow should support the biostatistician to analyse biological, genomic and clinical data in an efficient and effective way. Therefore the analysis tool should provide a structure of all available functions so that the user has the possibility to search for the appropriate function without losing much time. The biostatisticians have to know the appropriate statistical methods for the analysis. The algorithms should be self-descriptive to know which functionality is needed for the analysed data. They have to know the different databases which should be supported by the system in designing workflows for the analysis from biological and/or clinical questions. It should be possible for the biostatistician

and/or clinical questions - train the biologists/statistician involved in the experiment/study to analyse the data if feasible - analyse genomic data in the context of laboratory experiments and clinical data sets - discuss the analysed results with the biologists who produced the data and/or the clinicians who collected the samples - contribute to the design of the validation of the results - merge information from different data sets and perform meta-analyses to build robust models - merge the information from genomic analyses to other biological and clinical data, and to data coming from public databases or published material	Controllability Suitability for the task	to merge information from different data sets and perform meta-analysis to build robust models. Information from genomic analyses should be merged in an effective way to other biological and clinical data, and to data coming from public databases or published material.
Assumptions Francesca and Thierry have experience in programming with different languages and they know enough of statistics to understand the algorithms to contribute to their development and assessment. They are able to present complex information in an understandable way by the various communities of end-users involved. (Clinicians require some information presented in a different way than statisticians.) Modern biomedical research requires a large number of competences, from classical and molecular biology to advanced statistical methods. They need collaborative ways to work. The pre-existing clinical and biological knowledge, as for	Controllability Suitability for the task	The system should support the biostatistician in the different languages and algorithms to contribute to their development and assessment. Complex information must be presented whether for a physician or statistician in a consistent and understandable way. The statisticians need collaborative ways to work together because of modern biomedical research which requires a large number of competences from classical and molecular biology to advanced statistical methods. The binding to public, clinical and genomic databases must be possible.

instance from publicly available databases, is combined with new experimental results using statistical and bioinformatics methods. In a typical experiment, there is a disease with special parameters which must be analysed and the findings typically are validated by the biologists. From a global perspective, there is no systematic way to approach a biological experiment. The tools used depend on the way the question is asked by the biologist or by the clinician. From the bioinformatics / biostatistician point of view the biostatistician have a large number of tools available, and they pick the most appropriate ones based on the question of the researcher, usually after a long discussion. The biostatisticians claim that the following experience is needed: - in programming, - to know statistical methods, - to know genomic databases and concepts for a specific job within bioinformatics. It depends on the project, if it is a - laboratory trial, - clinical trial or - method driven project which event has to be taken. In the first two cases decisions are taken based on discussions with biologists and clinicians. In the third case decisions are taken after discussion with colleagues (other bioinformaticians or biostatisticians involved in the project).	Suitability for the task Controllability Self-descriptiveness Suitability for the task Controllability Suitability for the task Controllability	It must be self-descriptive and clear to the user which tools have to be used for the analysis process. There must be a clear understanding about the task; is it a laboratory trial, a clinical trial or a method driven project? The different types should be supported by the system. The access of computers have to be fast and smooth to public databases and web services, specific tools vary depending on the analysis and are in continuous development.

Powerful computers are needed, fast and smooth access to public databases and web services is desirable, specific tools vary depending on the analysis and are in continuous development.		
Routine activities The biostatisticians are working with clinical data, laboratory data or data from databases. They either analyse them or advice on the analysis, or use them to develop and assess methodologies for the analysis. The parameters vary from trial to trial, depending on the aims of the trial. The biostatisticians receive the data from biologists, clinicians, genomic facilities or public repositories. The form of the data varies, it can be tab-delimited text or other sort of text files from data bases. A typical biological or clinical analysis scenario consists of the following steps: 1. A researcher contacts the biostatistics group to ask for support. A preliminary meeting occurs during which the global scope of the work is defined. The analysis work is assigned to an analyst (bioinformatician/biostatistician) in the group. This usually is the most experienced person, or one that interested in conducting the analysis for the specific question. 2. A close working collaboration is initiated between the researcher and the bioinformatician/biostatistician. A	Suitability for the task Controllability Suitability for the task Suitability for the task	The biostatistician should be supported in choosing the right analysis in order to develop and assess methodologies or provide advice in a comfortable way. The data sets are varying and must be understandable for the statistician so that he/she needs no more information about it. The detailed analysis goals for the available data must be defined in a clear understandable way.

meeting takes place, during which a review of the available data and detailed analysis goals are defined.		
3. The data files are provided by the researcher to the bioinformatician. These files are typically Excel spreadsheets or text files describing the parameters of the experiment (e.g. patient age, blood pressure, treatment used, etc.). The researcher also has to provide the microarray files associated to the samples described in the previous files. In practice, it is usually the laboratory which has hybridized the microarrays which provide them directly to the biostatistician.	Suitability for the task	The system should ensure smooth data/files transfer between users.
4. A curation of the information in the files is done by the biostatistician: Inconsistencies between identifiers have to be removed (typical inconsistencies include character case, spacing and use of special characters in file names, etc…). Information in files has to be transformed into information usable by statistical software (e.g., information coded as colours in Excel files has to be transformed into textual labels), information available as Word documents with fancy formatting has to be transformed into tab-delimited files with specific format, etc… (This curation work can last for weeks until all inconsistencies are removed: e.g. preliminary analysis can identify incorrectly annotated samples, requiring going back to the original files to trace the mislabelling, etc…)	Suitability for the task Controllability	The user should be supported when inconsistencies of information in files exist. Transformation of information in files into information usable by statistical software should be supported in an efficient way.
5. The data files are loaded into R which is a scripting language (or other bioinformatics/biostatistics software) and are analysed in accordance to the discussion with the researcher. Intermediate findings	Suitability for the task	The user should be supported in an efficient and effective way when analysing the data sets. It should be possible for the user to look

are discussed jointly between the researcher and the biostatistician, possibly suggesting new directions of analysis or new validation experiments. This iterative process can last from a few days to several months, depending on the complexity of the questions under scrutiny. In parallel to the statistical data analysis *per se*, literature mining (usually based on online resources such as PubMed) is conducted to find alternative source of information on the problem. 6. For experiments in which interesting findings are done, a publication is prepared jointly with the researcher.		into literature (based on online resources such as PubMed to find alternative source of information for the existing problem.
Annotation quality varies depending on the source of the dataset, but usually minimal annotation is present. However, to fully understand the data discussion with the biologists and the clinicians more annotations are needed. This is true for example when the data is produced using new technologies for which standards are not yet present.	Self-descriptiveness	Annotation is a helpful amendment of the data sets. To reduce long discussions with biologists and clinicians annotation should be extended in a clear and understandable way.
Files usually require reformatting to be compatible with high-performance computational environments (e.g. Excel files cannot be loaded in a straightforward manner into R under Linux). Various tools can then be used to analyse the data. Typically, data can be imported in R and can be analysed using the packages available in R.	Controllability	One should have the possibility to transform files so that they are readable under different environments / operating systems. E.g., Excel files cannot be loaded in a straightforward manner into R. A transformation from an Excel file to an R file must be implemented so that the file is readable for R. After analysis the R file has to be transformed back into an Excel file to be able to send it to the clinician.
Part of these scripts can be re-used in the context of different analysis but they sometimes have to be adapted to the new set of data. This is basically the core of our work in the pure data analysis.	Suitability for the task	
	Suitability for the task	Every written script should be saved in an

How to write and manage scripts depends on a personal choice. Francesca tends to pass scripts to other people after she wrote them, to be used by them. This means that with some training depending either on the complexity of the tasks and methods or on the background of the student the scripts can be modified by him-/herself. This is true if one always deals with the same kind of problem, such as cancer in humans with a given family of microarrays, the re-usability is high. In broader biomedical research this may not be the case when other organisms are involved; for instance databases containing information about human genes have sometimes a different format than databases for mice or rats. Things tend to converge, though. In the context of a given established technology, such as Affymetrix microarrays, the quality control (QC) procedure of the raw data can be largely automated, although ultimately the results of the QC procedure could be at least in part automated. Repeated analyses can be imagined in the context of experiments in which new data is continuously produced and have to be incorporated with previously obtained results. However in the context of the bioinformatics core facility in Lausanne those represent only a small portion of the experiments. Repeatability of analyses has, however, to be guaranteed to ensure that findings can be retrieved. An important case requiring the repeatability is the production of figures for publications which might have to be adapted after the review	Suitability for the task Suitability for the task Suitability for the task Suitability for the task	efficient way so that it can be reused later on to repeat the analysis at a later time. The context of the different analysis of scripts must be clear and understandable, to decide which kind of script it is. The system should guarantee repeatability of analyses to ensure that findings can be retrieved. An important case requiring the repeatability is the production of figures for publications which might have to be adapted after the review process. In the context of analyses conducted with R, some packages wrapping complex

process. In the context of analyses conducted with R, some packages wrapping complex analysis pipelines in a single function call can be used. The user can generally control the most important parameters of such pipelines through arguments to the function calls, which is usually desirable to fine-tune the analysis. In general, however, a lot of high-level biostatistical analyses (i.e., beyond raw data quality control) are complex enough to require specifically written analysis scripts. In the context of ACGT, R will be one component of workflows that will become available. R can interface with other languages. One can write scripts to do the analysis in such a way that it can be reused to repeat the analysis later. For example, one may have generated a plot describing some findings in a research project; two years later new results may become available that the researcher wished to add to the same plot. It is thus important to preserve the procedure to generate the plot with a script. (Statisticians tend to call this “reproducible research”.) R is missing some features, for example, interactivity, which is a critical feature. For instance, one cannot interactively visualize large clustering results (e.g., heat maps) with zooming capability. Other interactive features are missing, such as menus. It would be nice to have a graphical environment where the user can use these functions with menus to click on a button to get the required information. Francesca often writes programs herself, but she re-uses a	Controllability Suitability for the task Suitability for the task Self-descriptiveness	analysis pipelines in a single function call can be used. The user can generally control the most important parameters of such pipelines through arguments to the function calls, which is usually desirable to fine-tune the analysis. A generated plot should be saved, too. If the plot has become important at a later time and the user wishes to add something to the same plot, it is necessary to preserve the procedure to generate the plot with the script. It would be necessary to support a graphical user interface with menus to click on a button to get the required information in an understandable and clear way. Or to get a mouse documentation if the button is mouse sensitive. The buttons of the system should be self-descriptive and easy to handle. Heat maps should be supported and the

lot of scripts or programs already available. If she has a heat map, e.g. with a clustering of patients' data, she can not zoom in or move the mouse over it and click on some items. It is extremely tricky. (It depends on the version of the software.) There are some visualization tools which are commercial and there are many which are free as well! A good visualization of the results of the analysis is compulsory. Every step that she has done in a graphical way should be memorized. This can easily be done in R, for example as she can track all her random activities. This is a way she scripts what she has done. She saves the commands to enable her to re-execute them later. It is a "walk through" for the whole analysis. It would be important to provide a mechanism to "replay" the user actions such that a specific graphics can be reproduced (e.g. addition of legends, change of axes, lines and points styles, change of viewpoint in 3D plots, etc…). Currently, the biostatisticians do it in R in a plain way without using interactive graphical interface, and only static plots can be produced. The biostatisticians wish to do it with workflows which are more powerful. But of course they do want more powerful tools all the time. When trying to address one broad question in a complex	Suitability for the task Suitability for the task Suitability for the task Suitability for the task	graphical presentation should be in a clear and understandable way. All available visualization tools should be supported by the system in an efficient and satisfactory way. . All interactive actions during the analysis process should be recorded to be able to replay them in a scripted way later on. This would make the work efficient and save time. This interactivity in the data mining tool is essential. All steps of analysis should be done in more intuitive way. To save the results and use them in another context again. The parallel work of different biostatisticians using the same normalized gene-expression matrix and demographics files must be supported in an efficient and

biological or clinical experiment many derived questions are usually raised. E.g., in a gene-expression experiment investigating the response of patients to a drug, one might want to investigate the role of specific distinct biological pathways, whereas another one would like to stratify patients according to different markers, in order to investigate the response. These alternative investigations are sometimes conducted by different biostatisticians in parallel, using the same normalized gene-expression matrix and demographics files as starting point. Publication writing is another example of process in which several users work on the same object. In biostatistics only a single user (i.e., a biostatistician) usually works in a specific session. However, it is likely that advanced tools will permit many people working in parallel in the same session. There are specified working steps in routine analyses, such as in quality control. However most of the working time of biostatisticians is spent on non-routine work. Thus, although there are families of similar problems from the viewpoint of statistical analysis, the details of each analysis usually require a specific approach. Through the use of scripted languages, such as R, which are highly flexible, sound statistical analysis flows targeting very precisely the question under scrutiny can be designed. The research process is largely random, and requires tools which remain as flexible as possible. (This is why R, despite numerous flaws, remains largely used in the biomedical community.)	Suitability for the task Suitability for the task Suitability for the task	consistent way. (Version control for writing publications by several users should be supported, too. Not relevant for the Workflow) To support the different users efficiently in their work. Flexibility should be improved by providing advanced visualization capabilities.

In typical experiments, the core results of a biostatistical analysis are usually figures, plots and tables. (A table can contain other information than just numbers, e.g., gene or pathway names, patient identifiers, etc.) These raw results are usually easily reusable to continue the analysis in which they have been produced, or to extend other analyses (e.g., a gene signature identified in an analysis on breast cancer can be used in the context of another type of cancer to identify similarities or discrepancies). The raw results also form the basis of the textual description and interpretation which is built upon them (which form a more abstract family of results). In complex cases, a new approach (algorithm) to the analysis of data is required. Such new algorithms can then be used in the context of other studies with a similar structure. The feedback of the analysis takes the form of a discussion between the biostatistician and the researcher responsible of the trial. Once an agreement has been reached, the reactions of the research community are an important feedback on the work (reviewers' comments in journal publications, questions in conferences, etc…). The biostatisticians write Perl scripts. Many scripts are written. Sometimes Francesca receives data directly either from biologists or from clinicians, but rather from genomic facilities, database managers, or from public repositories.	Suitability for the task Suitability for the task Suitability for the task Suitability for the task Suitability for the task	Interactivity in graphics is a desirable feature. It would be important to provide a mechanism to "replay" the user actions such that a specific graphics can be reproduced (e.g., addition of legends, change of axes, lines and points styles, change of viewpoint in 3D plots, etc…). The core results of a biostatistical analysis are usually easily reusable to continue the analysis in which they have been produced, or to extend other analyses (e.g. a gene signature identified in an analysis on breast cancer can be used in the context of another type of cancer to identify similarities or discrepancies). This should be possible and conducted by the user in an efficient and understandable way. It should be possible to conduct a new approach (algorithm) to the analysis of data, so that this new algorithm can be used in the context of other trials having a similar structure. Perl scripts should be supported, too.

Before collaborating with them, she has to understand more about the clinical data or biological data used in the trial. She has an access to several databases to retrieve the data she needs for analysing. For the clinical data she usually does not do any data processing; this is done by a database manager and/or a medical statistician. For gene expression microarray data some processing and annotation is done in the genomic facilities producing the data, some is done by her using freely available resources. She also receives data which is usually already processed and ready to analyse. In contrast to Thierry, Francesca usually has some visualization and pre-processing of the data done by the facility producing the data or the biologists producing the data. Sometimes this pre-processing needs to be re-done in order to be consistent with previous analyses. Francesca's clinical component is that she analyses genomic data usually in the clinical contest. She correlates the biological data to the clinical data. She uses several available tools, e.g. tools for clinical databases, genomic data, proteomic data, ontology tools, etc. Sometimes she needs to develop some of these tools herself. She does not have ideal visualization for all these tools and sometimes the passage from a tool to another is not smooth/automatic.	Suitability for the task Suitability for the task Suitability for the task Controllability	The access to several databases must be given in an easy and understandable way. The user must have the possibility to save scripts efficiently to reuse them later. The user should be supported while processing and annotating the data sets. Correlation of different data sets should be supported in an efficient and effective way. All available tools should be supported by the system. The interface should be self-descriptive and easy to handle. The system should support the user to give him/her action guided information, so that he/she knows in every situation what to do next. The switch from one tool to another should be done in a smooth and easy way. Consistency of visualization and all pre-processing of the data are required to be able to combine it with previous analyses. Analyses of genomic data and the correlation with biological data in a later step should be done in an efficient and

		effective way Visualization of all these different tools should be supported so that the switch from one tool to the other can be done in an efficient and easy way.
Special features during the working process Most of the procedures are not fully automated and standardised today. ACGT will deliver a platform where automated workflows for analysis of user designed workflows can be executed. With all these data the bio molecular researcher can make an analysis and give the analysed data back to the clinicians, or vice versa. This should be reached in ACGT to share and compare appropriate trials and make the analysis by the bio molecular researcher much easier and efficient. It is advisable to conduct the development of analysis procedures without interruption otherwise one might lose the logic of the data analysis flow. Important is to transfer the analysed data to colleagues or save the intermediate state of analysis. So the biostatistician can use it again or continue without doing every step once more. To freeze the state of script and use it again. In principle, the statisticians have to register the version of the analysis tools used because if results differ this might	Suitability for the task Suitability for the task Error tolerance Suitability for the task	Designing and generating workflows should be conducted efficiently and smoothly. All data sets the molecular researcher gets from the clinician are analysed efficiently with the support of the system. The data flow should be self-descriptive. Interrupted flows should be continued efficiently without losing the logic of the analysis. Therefore all data flows should be saved with some description to reuse them at a later time. When an error occurs it must be easy to correct the mistake efficiently. The error message of the system should be displayed in a clear and understandable way. The language of the user should be used. To transfer the analysed data or an intermediate state of it to colleagues the

help identifying the reason for the change (e.g., the algorithm associated to a function has changed or the annotation for genes has evolved to incorporate new knowledge).		data has to be saved efficiently and in a clear form so that the user can work with it or continue without doing every step once more. The version of the analysis must be automatically indicated to identify the algorithm used for the analysis as well as the annotation for genes which has evolved to incorporate new knowledge.
Organizational conditions Thierry prefers to use the Linux platform for the analysis of data sets. Francesca uses both Linux and Windows platform. Stress factors in the academic environment are mostly related to working conditions (salary, workspace, lack of long term contracts, etc.). In biomedical collaborations, a point of stress may be linked to the lack of understanding regarding how the others are working. For instance, when a bioinformatician requests some data from lab researchers, it is quite common to receive an Excel document with the useful information in a format not exploitable directly. A lot of extremely frustrating (and in principle unnecessary) reformatting efforts are required to make these data available to advanced analysis tools such as R.		The documentation / annotation of the requested Excel files do not exist or are too small to understand the whole context. To minimize or eliminate the reformatting efforts for a better understanding there must be a support by the user who creates the file or by the system itself.

Regarding software usage, factors of stress are: • strong differences in user interface between versions of the code • differences in code outputs for the same input when different versions of the code are used (e.g. if default parameters of an algorithm are changed) • lack of clear documentation (written in simple and correct English) • incomplete/obsolete documentation • unclear error messages • arbitrary change in code output, depending on unrelated changes in the input (e.g. in R, matrices converted to vectors depending on dimensions) • lack of ways to report bugs • lack of ways to have bugs corrected (e.g. bug reports ignored by developers) • lack of ways to have support (either from the community of users or from the developers) Punctual existence of publication or reporting deadlines may also bring a share of excitement. Francesca has to work with lots of tools, e.g. clinical databases, proteomic data, genomic data, and others. The main problem is that these tools do not communicate easily with each other. Some of these tools have a fantastic visualization. But the problem is not the missing visualization. It is the practice of these tools, to switch between different visualizations in one	Controllability Self-descriptiveness Error tolerance Suitability for the task	Stress factors should be reduced severely: • to make the versions compatible; a version control of the data sets • different code output has to be recognized by the system • full documentation in a correct and simple English is required • to check incomplete / obsolete documentation • error messages have to be clear and understandable for the user • support for arbitrary changes in code output depending on unrelated changes in the input • the user should have the possibility to report bugs and get the correction effectively • developers should recognize bug reports and correct them in an efficient way • the user should have support from the community of users or from the developers for knowledge transfer The user needs a better communication of the analysis tools among each other. The problem of these tools is that the user

environment. E.g., to switch from a classical heat map to the information of the patient and to the information of the trial. In a classical trial the clinician provides the data on the patient demographics, on the treatment, other clinical data, clinical markers and so on. From the same patient Francesca has genomic and proteomic data on the same samples. She puts these data sets together using the R tool (or other tools) and processes them. Her intention is to work in one environment with all these heterogeneous data sets and switch from one data to the other without changing the environment. To put different data sets together is error-prone. It takes a lot of time to write the data in a proper way.	Suitability for the task	can not switch between different visualizations in one environment. This should be realized in a smooth and efficient way. E.g., to switch from a classical heat map to the information of the patient and to the information of the trial The user has to work with lots of tools which should be supplied in one environment. He/she wants to combine all the heterogeneous data sets of the different tools and process them with R without switching from one environment to the other. This should be possible with the new tool / implementation. The work should be conducted in an efficient and effective way. There should be a print button, too, to discuss results on paper with colleagues for better understanding.
Other comments to critical incidents which already occurred A major drawback in the use of R is the inconsistent specification of the defaults for the functions between versions of the software (or, for a given version, between different platforms supporting the code). For instance, while the default character to specify comments to read in text files was the hash sign in some early versions of the code, this was changed to have no default comment	Self-descriptiveness Suitability for the task	Inconsistent specification of the defaults for the functions between versions of the software (or, for a given version, between different platforms supporting the code) is a great problem which should be solved.

sign. Another example is the fact that a column containing integers was read as “factors” in a Windows version while it was read as “integers” in the corresponding Linux version, with the critical effect that array elements were ordered differently in the two data sets.		