# Project scope and objectives

## Background

As a bioinformatics marketplace, Clarity makes use of computational resources such as the Centre for High Performance Computing (CHPC). This computing cluster in turn uses a scheduler (PBSpro) to administrate its resources based on the amount of time required for each job.

There is a tendency to overestimate required resources when creating a job, mainly to ensure sufficient allocation for what is required. However, PBSpro tries to fit each job into the total capabilities of the system. The bigger the pieces, the more unlikely a fit will be found timeously. Overestimating resources therefore extends the amount of time it will take for a job to start. This is especially problematic with the “bigmem” queue, as the limitations on the number of jobs running and queued are very stringent.

Additionally, functions of the resources and how to use them efficiently is often not well known by the users. For instance, certain processes might be better handled in parallel, thus using fewer resources altogether and allowing for a job to start and be executed quicker than standard, non-parallel routes.

## Aims

1. Establish protocols for running parallel jobs.
   1. Understand principles of parallel computing.
   2. Perform scaling analysis on eligible programs.
   3. Produce standard operating procedure to perform scaling analysis.
2. Determine optimal resources for programs on the CHPC that are not fit for parallel runs - bigmem.
   1. Initiate a dynamic list of programs repeatedly used my members of the CLARITY team.
   2. Perform tests on the programs to determine the resources required to run a job optimally (maximize resources as well as time for scheduler to run job).
   3. Document steps to perform the analysis above so that others can repeat it.

# Materials and Methods

## Literature review

An internet search has been performed and found that the subject documentation is either lacking or composed with experienced programmers in mind. Understanding these documents therefore will require more time and research to be understood by individuals with a history in biology education. The information will then be collated and presented in a way accessible to bioinformaticians who do not have any formal programming background. The document will also be made available to all members of CLARITY and Diplomics for review and will be kept up to date by the members.

## Parallel computing

With this knowledge, the parallel computing practice will begin. The CHPC has documentation available on scaling and his will be used to perform such a scaling analysis using the FastQC program and Oxford Nanopore Technologies (ONT) data. During the course of this analysis, an SOP will be written. The idea is that scaling analyses be made available to other members of the CLARITY team so that they do not need to repeat it. Where scaling analyses are missing, members will be encouraged to use the SOP and add their analysis to the living document. The parallel computing knowledgebase, which includes the information document, scaling analyses of programs and scaling analysis SOP, will be accessible on GitHub to members.

## Bigmem

A similar exercise will be performed on programs that cannot be run in parallel but require use of large amounts of memory. In theory, scaling can also be done for these projects, as the required resources should still increase with the size of the input and/or output. The ONT data is usable here again, but this time, the data will be aligned to a reference. The first step will be to align all the human data to the human reference genome, and this will serve two purposes. The first purpose will be to serve as the “blind approach”, i.e. without knowing the optimal resources and likely overestimating them, then noting how long it takes PBSpro to start the job. The second purpose will be to use the resulting alignment file to generate smaller pieces of the input data and genome, thereby allowing “scaling” analysis. Finally, the whole dataset will again be aligned to human reference genome using optimal resources calculated from the analysis. The time it takes for the job to start will be compared to the first “blind” job submission as well as all of the “blind” submissions during the scaling analysis to determine whether resource tailoring saves time.

During the course of the above analysis, an SOP will be written to be made available. As with parallel computing, a knowledgebase will be made available on GitHub consisting of similar components.

# Project admin

## Outputs

A reference document will be produced serve all CLARITY personnel (obj. 1a).

Scaling analyses will be made available to CLARITY personnel to inform their job submissions (obj. 1b).

A reference document will be initiated to allow analysts to calculate the optimal computing resources based on the input data (aim 2).

A standard operating procedure will be written so that if a program is not in the above reference document, the CLARITY analyst can perform the calculations themselves in an efficient manner and subsequently, add their results to the program reference list (obj. 2c).

The client will maintain ownership of the data as well as gain ownership of any outputs produced.

## Client responsibilities and deliverables

The client, Diplomics/CLARITY will provide the data (described above) and the computing resources.

## Milestones and timeline

The three aims are independent and can therefore be progressed simultaneously.

|  |  |  |  |
| --- | --- | --- | --- |
| Category/Aim | Objective | Working hours | Projected completion |
| Admin | Analytical study plan | 4 | 1 May 2024 |
| Resource estimation | 4 | 1 May 2024 |
| Write report | 16 | 17 July 2024 |
| Buffer\* | 34 | 31 July 2024 |
| Aim 1 | Understand parallel computing | 40 | 5 June 2024 |
| Scaling analysis | 10 | 26 June 2024 |
| Scaling knowledgebase | 8 | 3 July 2024 |
| Aim 2 | Align all data | 2 | 15 May 2024 |
| Break data into smaller pieces | 8 | 22 May 2024 |
| Resource scale analysis | 16 | 5 June 2024 |
| Bigmem knowledgebase | 8 | 12 June 2024 |
| Total |  | 150 | 31 July 2024 |

\* Adjusted hours and completion date so to ensure that true hours and completion date of project does not exceed budgeted amount.