Human Genome Project[1]

By Jérémie Gallien[2] and Scott A. Rosenberg[3]

Scott was now a couple of weeks into his MIT Leaders for Manufacturing program internship at the Whitehead Institute in Cambridge. The exciting premise for his project was that the genome finishing group at work there could benefit from a sound flow analysis of the type usually applied in manufacturing environments. While he had received great support from his supervisors all along and felt that he had already acquired a reasonably good handle of the work in his area of scope, he also knew that he still had to prove his worth: in an environment with scores of PhDs, many world-renown scientists and even Nobel prize winners walking through now and then, nobody was going to settle for small talk and unsupported recommendations.

Indeed, the sequence finishing operation seemed considerably more repetitive and process-oriented than anything else at the Whitehead, an institution known for breaking new grounds in Biology through scientific experiments never attempted before. Because of the need to accurately forecast the final completion date of the genome and to plan for staffing levels, the largest complaint of the scientists overseeing the finishing group was by far the variability of weekly output. Scott suspected that the practice of bundling many tasks into a single project assigned to each finisher and the informal, on-demand policy followed when assigning new projects played no small role in this variability. What he did not know yet however was how to quantify these effects so he could convince the Whitehead managers to change their procedures.

The Human Genome Project

Arguably the most important undertaking in life sciences since the discovery of DNA in 1953, the Human Genome Project (HGP) began in 1990. Its simple but ambitious goal is to sequence the entire genetic makeup of the human species, which will enable decades of evolutionary and medical research on cross-genomic comparison, disease risk detection, gene therapy and possibly uncountable other applications not even imagined yet by scientists. Primary responsibility for the sequencing fell to large genome centers like the Whitehead Institute at MIT, Washington University, Baylor University, and the Sanger Center in Great Britain, with dozens of smaller centers around the world also contributing.

[1] This version: July 2003.

[2] Corresponding author: MIT Sloan School of Management, Cambridge, MA 02142

[3] Most material in this case originates from Scott Rosenberg's Leaders for Manufacturing Master's thesis, "Managing a Data Analysis Production Line: An Example from the Whitehead/MIT Center for Genomic Research," Sloan School of Management and Dept. of EECS, Massachusetts Institute of Technology, June 2003.

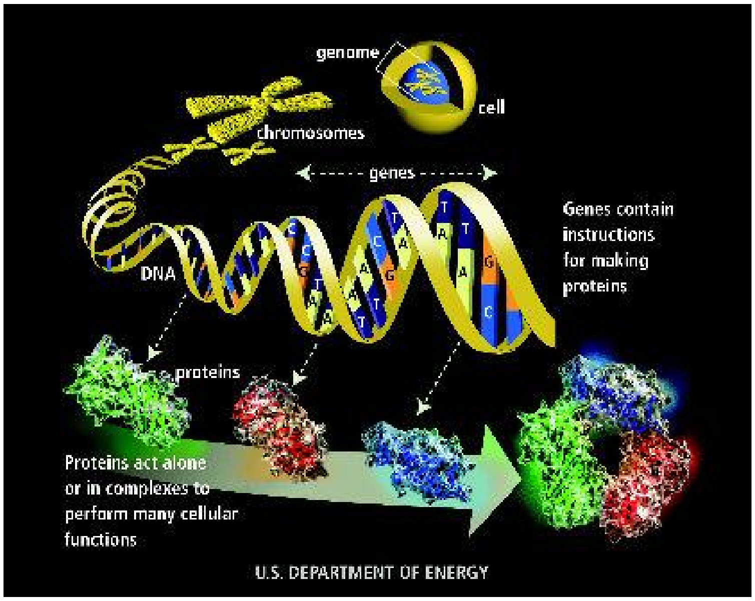
In 2000, a draft sequence of the human genome was published. However, it contained many absent, ambiguous, or conflicting regions of DNA. In the time since the draft's publication, genome centers like the Whitehead have concentrated their energies on systematically clarifying these problematic regions. This process is called *finishing*, and is both the current bottleneck of the genome project and the focus of Scott's work.

DNA Sequencing Background

Deoxyribonucleic acid (DNA) is the genetic building block upon which all known life regulates its daily function and long-term evolution. Constituting the *chromosomes* found in the nucleus of human cells, DNA is itself comprised of long strings of just four nucleotide bases called adenine (A), guanine (G), cytosine (C), and thymine (T). Active sequences of DNA that are hundreds or thousands of base pairs long, called *genes*, are translated into *proteins* during the course of cell activity. Proteins, in turn, enable all of life's most basic functions.

Structurally speaking, DNA is a stable polymer that arranges itself into a double helical structure as shown in Figure 1. Long sequences of nucleotide bases form one half of the structure. Each base also bonds to its complementary base in the other half of the structure: A pairs with T and G pairs with C. Thus, a sequence of "ATTGC" bonds to its complementary sequence "TAACG". All told, the human genome consists of more than three billion DNA base pairs and an estimated 30,000 genes.

Figure 1. Relationship between cells, chromosomes, DNA, and proteins.[4]

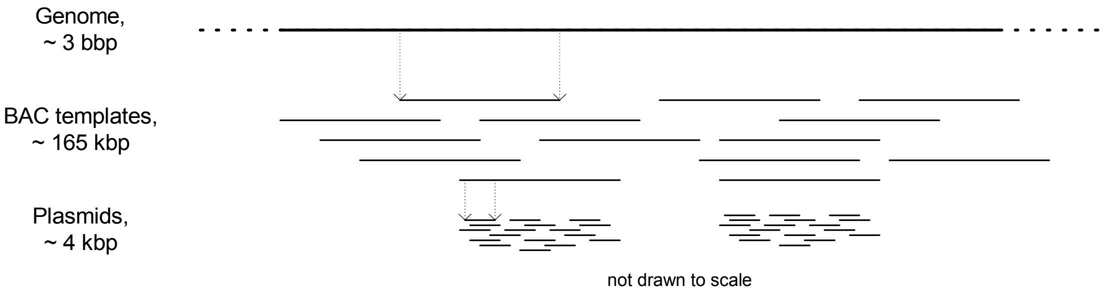

genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

Today's state-of-the-art gene sequencing technology proceeds by breaking large DNA samples into small segments, determining the exact DNA sequence of those small segments, then reconstructing sequence from these segments into a composite view of

[4] Source: U.S. Department of Energy, http://www.ornl.gov/TechResources/Human_Genome/publicat/primer2001/1.html.

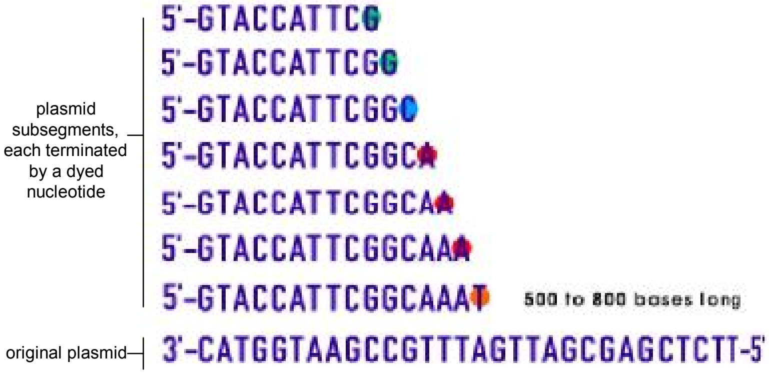
the original sample: DNA donated by a small set of consenting, anonymous individuals is first purified, then enzymes are used to break it down into smaller segments; From the mix that results, segments with a length of approximately 165,000 base pairs (165kbp), called *BAC templates*, are isolated. An engineered version of the bacteria *E. coli* can then be tricked into carrying and reproducing this human genetic material millions of times in just hours. To accomplish a further reduction in sample size necessary to direct sequencing, BACs are then sheared through a physical process and filtered, producing DNA segments of uniform size, usually between 4kbp and 10kbp. Once isolated, each such segment becomes known as a *plasmid* (see Figure 2).

Figure 2. Two phase break-down of genome into BACs and then plasmids.



After amplification through *E. coli*, plasmids are placed in a solution containing special DNA base pairs that are tagged with a fluorescent dye. By raising the temperature of the solution, the plasmid DNA, which normally resides in a paired helical structure, can be induced to separate. When the temperature is lowered, an enzyme in the solution reconstructs the helical structure by grabbing base pairs from the surrounding solution. Whenever the enzyme selects a dyed base pair, however, the reconstruction process stops, leaving a DNA segment that is prematurely terminated by a dyed A, T, G, or C. By cycling the heat many times, technicians can produce a solution containing a wide array of dye-terminated segments of various sizes (see Figure 3).

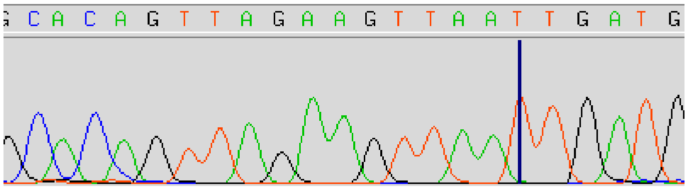
Figure 3. Plasmid dye-tagging process. [5]



[5] This graphic is from an animated educational toolkit provided by the National Human Genome Research Institute, http://www.genome.gov/Pages/Education/Kit/main.cfm.

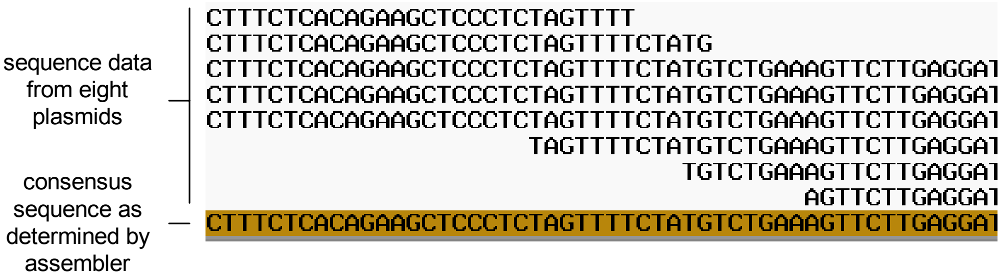
This solution becomes the input to *detection*, the last of the laboratory stages of the gene sequencing process. The solution of dyed plasmid segments is placed at one end of a long capillary. A charge causes the DNA to migrate through this capillary, with smaller segments racing ahead of larger segments because of their lighter molecular weight. At the end of the capillary, where the segments gradually emerge, a laser illuminates the dyed base pairs at the end of the DNA molecules. A sensor detects the continuously varying illumination and records it in a data file. A piece of software then analyzes this data and makes a base-pair determination (see Figure 4).

Figure 4. Example output from detection process.



What remains of the gene sequencing process is strictly information processing. A software tool called an *assembler* attempts to match plasmids from a BAC by similarities in their sequence. If everything works correctly, the assembler will reconstruct a single view of a BAC's underlying sequence (see Figure 5).

Figure 5. Example assembly with eight reads and consensus.


sequence data from eight plasmids
consensus sequence as determined by assembler

```
CTTTCTCACAGAAGCTCCCTCTAGTTTT
CTTTCTCACAGAAGCTCCCTCTAGTTTTCTATG
CTTTCTCACAGAAGCTCCCTCTAGTTTTCTATGTCTGAAAGTTCTTGAGGAT
CTTTCTCACAGAAGCTCCCTCTAGTTTTCTATGTCTGAAAGTTCTTGAGGAT
CTTTCTCACAGAAGCTCCCTCTAGTTTTCTATGTCTGAAAGTTCTTGAGGAT
                     TAGTTTTCTATGTCTGAAAGTTCTTGAGGAT
                               TGTCTGAAAGTTCTTGAGGAT
                                       AGTTCTTGAGGAT
CTTTCTCACAGAAGCTCCCTCTAGTTTTCTATGTCTGAAAGTTCTTGAGGAT
```

Unfortunately, in many cases the assembler fails to construct a complete rendition of the BAC, leaving regions known as *gaps* where the DNA sequence data is of poor quality or missing. This is where the finishing group intervenes.

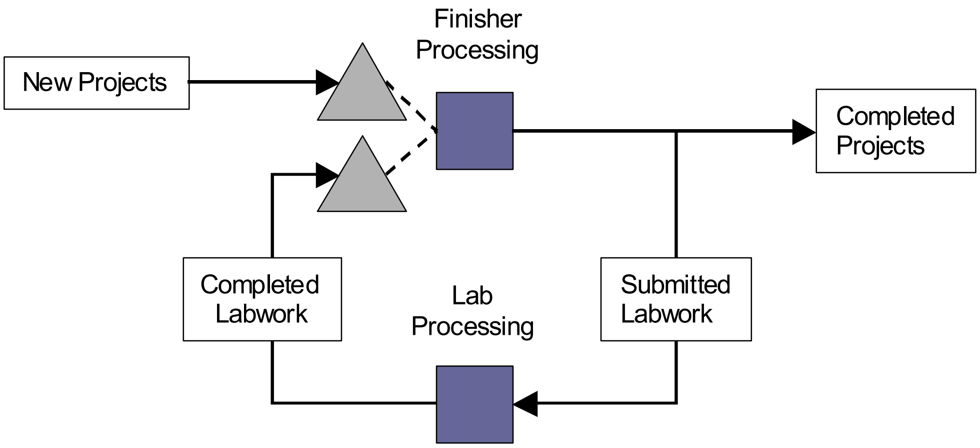
The Finishing Group

Closing gaps is the primary function of the finishing group of the Whitehead Institute, employing at times as many as forty human analysts. In some cases, finishers can close a gap by editing the data already present in the BAC assembly using various specialized software. In other cases, finishers must select and order laboratory procedures, then analyze them in order to discover missing sequence information. In classic manufacturing terms, finishing represents the inspection, quality assurance, and rework phases of the gene sequencing process.

There are many possible reasons why gaps form; in addition to operator errors examples include toxicity of plasmid DNA to *E. coli* and DNA resistance to certain chemicals used. Very few of these reasons are fully understood however, as it is often extremely difficult or economically infeasible to determine how a particular gap arose. Generally speaking, finishers triage two types of gaps. A *captured gap* is spanned by genetic material from a single plasmid. The finisher may be able to perform a lab procedure on the plasmid in order to discover the missing sequence. An *uncaptured gap*, on the other hand, occurs between the sequences of two or more plasmids. Thus, the assembler has no basis for joining the sequence on either side of the gap. The finisher must then use other, more complicated techniques to discover the missing DNA sequence. For a variety of reasons, uncaptured gaps usually prove more difficult than captured gaps.

Finishers often find themselves in a catch-22: to select an appropriate laboratory procedure, they must understand the underlying sequence, but the sequence is missing. In practice, they must make educated guesses about the underlying sequence and the likelihood that various laboratory techniques will succeed. Their decision is influenced by the condition of the DNA near the gap; it is also influenced by the ability of their informatics tools to highlight those conditions. Most importantly, finishers' decisions are guided by their skill and experience: whereas some experienced finishers may be able to close a gap based on the information already present in an assembly, less experienced finishers may feel they need laboratory work. In any case, finishers are often unable to close a gap after a single round of additional laboratory tests (a *work cycle*) and must try again until they succeed in closing the gap. Each attempt generates information that offers new insights into how the finisher should proceed. For example, the gap may have been partially closed, indicating that the previous procedure worked, albeit incrementally. Alternatively, a failure may indicate that the underlying DNA is resistant to the chosen procedure. In still other cases, the procedures may fail uniformly, raising the possibility that the lab committed an error. With the information they gain at each attempt, finishers proceed in a trial-and-error, iterative fashion until they succeed in closing the gap (see Figure 6 for a process flow diagram, and Exhibit 1 for data on gap closure probability and finisher processing times).

Figure 6. Process flow diagram for a finisher's work.



There are both delay and cost implications to the iterative nature of this workflow: projects submitted to the lab are only returned after an average of 70 hours, with a standard deviation of 10 hours (while the lab is shared across all finishers, it may be assumed for the purpose of analysis that the lab has infinite capacity). In addition, each lab iteration is estimated to cost about $200.

In part to promote a sense of work ownership, Whitehead has historically converted each BAC assembly into a finishing project and then assigned it to a single finisher until project completion. BACs however contain a variable number of gaps, so the volume of finishing work associated with each project assignment may vary significantly (see Exhibit 2 for data on the number of gaps per BAC assembly project). When a finisher starts working on a project (either a new project or from the completed labwork queue), he/she first works on all its gaps, then (if appropriate) sends it to the lab as a whole for further tests on all the gaps still unresolved at that point. A project is completed only when all its gaps are closed.

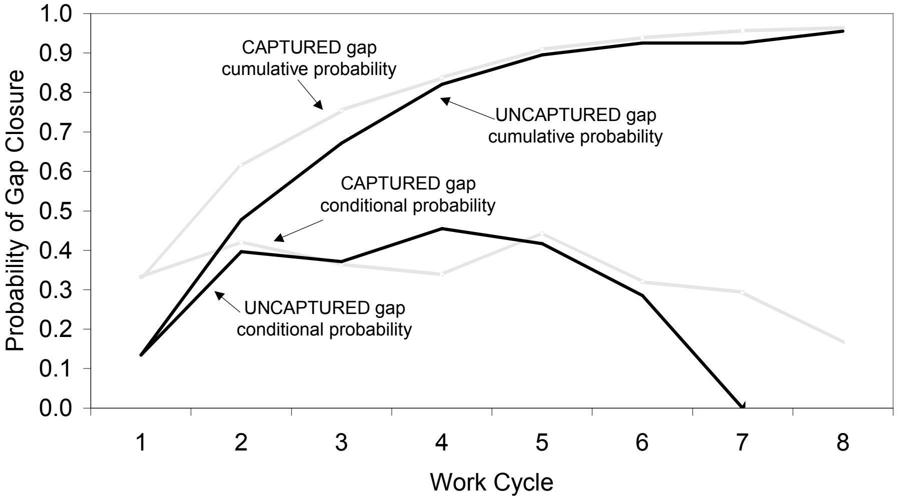
The assignment of new projects to finishers is done fairly informally. For the most part, managers have been following an “on-demand” work release policy whereby finishers independently request new projects to be assigned to them. In the current phase of the HGP, requests are automatically granted as the number of new projects available seems virtually infinite. For most finishers, the primary consideration driving a new project request is the immediate concern of running out of work. Many also acknowledge the objective to constitute a portfolio of both hard and easy projects as part of their work-in-in progress, so they can more easily meet the short-term production goals sometimes set by managers. In fact, a common pattern is for finishers to request (and obtain) a new project whenever their queue of completed lab work projects falls below a “watermark” level of 4 to 5 projects. Finishers usually give priority to new projects over existing ones that came back from the lab.

Exhibit 1: Finisher Performance

Gap Closure Probability

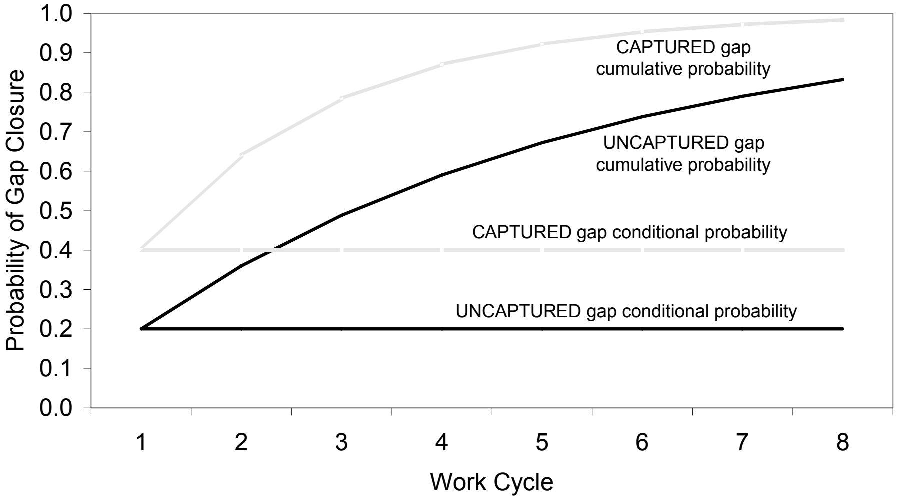
Figure 1 below shows the observed probability of gap closure by work cycle for Whitehead's finishers in 2002. The *conditional probability* lines represent the chances that a gap is closed in its N^th^ work cycle given that it was not closed in the preceding N-1 cycles. The *cumulative probability* lines represent the chances that a gap is closed by the end of the N^th^ work cycle.

Figure 1. Observed gap-closing probabilities at Whitehead.



Because of biases in collection of the data shown above and to simplify analysis, it may be assumed that captured (resp. uncaptured) gaps close with probability 0.4 (resp. 0.2) at each and every cycle, as in Figure 2:

Figure 2. Idealized gap-closing probabilities.



Time per Gap per Cycle

In practice, the time that analysts spend on each gap in each work cycle varies with each analyst, gap and cycle. For the purpose of analysis however it may be assumed that an average finisher spends exactly 1h per captured gap per cycle, and 1h30 per uncaptured gap per cycle. Also, analysts may be assumed to effectively work 35h per week on closing gaps.

Summary of Finisher Model Data

The data that may be used for modeling purposes as described in this exhibit is summarized in Table 1:

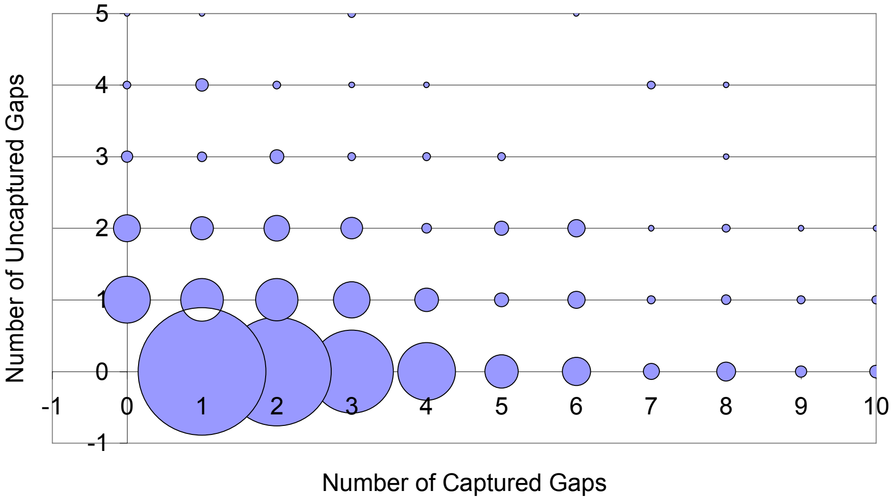
Table 1. Model of individual finisher performance.

Conditional gap closing probabilities	
Captured gap / cycle	0.4
Uncaptured gap / cycle	0.2
Average time / gap / cycle	
Captured gap (hrs)	1h
Uncaptured gap (hrs)	1h30
Effective Workweek (hrs)	35h

Exhibit 2: Gap Distribution per BAC Project

Because BACs are spliced out of a genome by means of an enzymatic process that is unaffected by the sequence problems that may lead to a gap, the incidence of gaps within BACs is fairly random. Figure 1 shows the observed frequency of projects according to their gap count in Whitehead's portion of the human genome. Balloon size indicates the relative frequency of a project with [x,y] captured and uncaptured gaps. Projects with zero gaps ([0,0]) are excluded from the distribution because they generally require little finishing work.

Figure 1. Observed gap distribution.

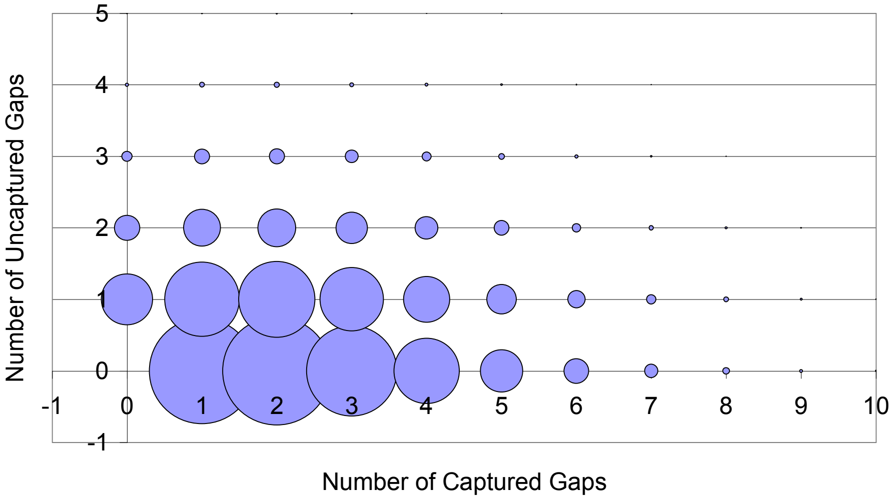


The mean rate of gap occurrence in Whitehead's projects is 2.1 captured gaps per project and 0.5 uncaptured gaps per project. While the actual data exhibits a slight correlation, for the purpose of analysis it may be assumed that captured and uncaptured gaps occur independently. Secondly, it may be assumed that the number of gaps in a project occur according to a Poisson process with the same observed means as the empirical data (i.e. 2.1 captured gaps per project and 0.5 uncaptured gaps per project). So if C and U denote the number of captured and uncaptured gaps per project, respectively, we have:

$$P(C = x, U = y) = \frac{\alpha^x}{x!} e^{-\alpha} \frac{\beta^y}{y!} e^{-\beta}, (x, y) \in \{0,1,2,...\},$$

with α = 2.1 and β = 0.5. Figure 2 shows gap distribution (excluding projects with no gaps [0,0]) according to this simplified model:

Figure 2. Idealized gap distribution.



A number of discrepancies with the empirical model are easily spotted. Gap counts fall off more precipitously in the theoretical model. Also, projects with one uncaptured gap appear more common than in the empirical data. However, in light of biases in collection of the data shown in Figure 1 and the resulting analysis simplification, the manager of the finishing group feels that this is an appropriate assumption.

Case Assignment for Human Genome Project

1. Consider a finisher just starting to work on a project with N remaining captured gaps and M remaining uncaptured gaps (the project may have gone through a number of cycles already):

   (i) What is the total time that the finisher will spend working on the project during this cycle?

   (ii) What is the probability distribution for the number of captured and uncaptured gaps still remaining when the finisher is done working on the project for this cycle?

   Hint: To model these features in Simul8, you may want to associate with each work item representing a project four number-valued labels keeping track of the both the initial and remaining number of captured and uncaptured gaps. The labels representing the remaining numbers of gaps can then be updated each time the project goes through a cycle through the probability distribution determined in (ii) above (in Simul8, labels can be used directly in the parameter fields defining probability distributions). In addition, you may also use a number-valued label representing the remaining finishing time requirement, and update it through the formula established in (i) whenever the number of remaining gaps change. When the project goes through the work center representing the finisher, the appropriate service time can then be obtained through a label-based distribution associated with this last label.

2. Build a simulation model with Simul8 to determine the average project completion time for a single finisher working under the on-demand work release policy with a watermark level of 4 projects (i.e., the finisher requests a new project whenever the total of projects in his/her Completed Labwork and New Projects queues falls below 4). What is the average and standard deviation for the number of captured and uncaptured gaps in the completed projects each week?

3. How are the results of question 3 modified when instead a just-in-time work release policy is followed? (i.e. new projects are only assigned to finishers when no work is available to them otherwise). How is the average level of WIP affected? Interpret your results.

4. Adapt your simulation model to compute the average and standard deviation of the number of captured and uncaptured gaps completed by one finisher in one week if all the projects had a single gap and the just-in time policy were followed. That is, now the unit of work is no longer a BAC but instead a single gap, but the proportion of captured and uncaptured gaps remains exactly the same as before. Interpret your results.