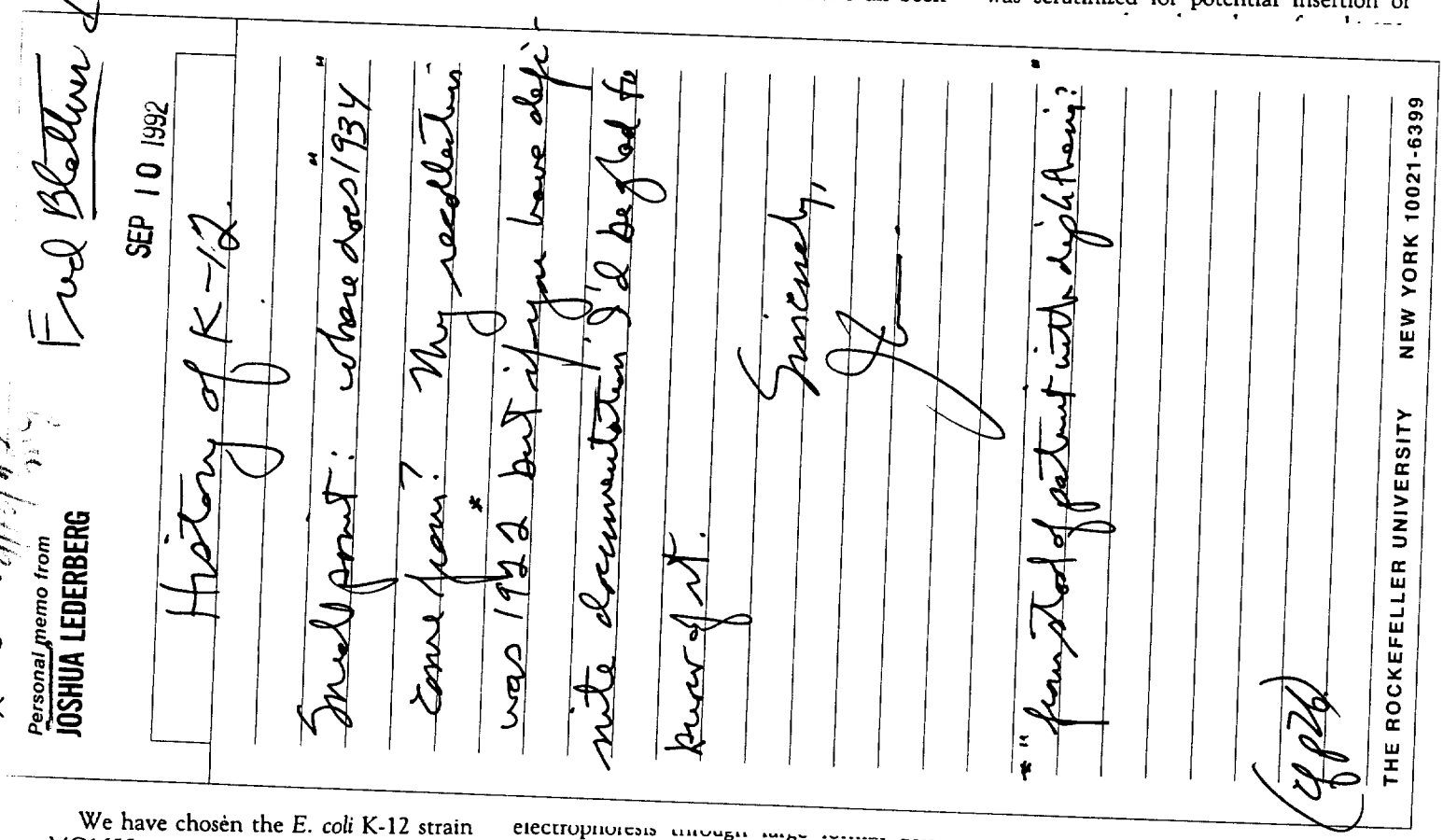
Analysis of the *Escherichia coli* Genome: DNA Sequence of the Region from 84.5 to 86.5 Minutes

Donna L. Daniels, Guy Plunkett III, Valerie Burland, Frederick R. Blattner

The DNA sequence of 91.4 kilobases of the *Escherichia coli* K-12 genome, spanning the region between *rrnC* at 84.5 minutes and *rrnA* at 86.5 minutes on the genetic map (85 to 87 percent on the physical map), is described. Analysis of this sequence identified 82 potential coding regions (open reading frames) covering 84 percent of the sequenced interval. The arrangement of these open reading frames, together with the consensus promoter sequences and terminator-like sequences found by computer searches, made it possible to assign them to proposed transcriptional units. More than half the open reading frames correlated with known genes or functions suggested by similarity to other sequences. Those remaining encode still unidentified proteins. The sequenced region also contains several RNA genes and two types of repeated sequence elements were found. Intergenic regions include three "gray holes," 0.6 to 0.8 kilobases, with no recognizable functions.

Complete genomic sequences, including those of viruses, plasmids, organelles, and [illegible] of lambda prophage and F factor without treatment by mutagens. Other common laboratory strains of *E. coli* have all been



We have chosen the *E. coli* K-12 strain MG1655 to represent the wild type for sequencing (9). It was derived from the original K-12 (isolated in 1934) by curing it

The authors are in the Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706.

relatively small team of technicians aided by student workers. At this point examination of the sequence data was limited to quality control checks. Ambiguities, where several determinations of an individual nucleotide (nt) differed (*12*), averaged about 1 per 100 nt.

A second team, working with computer assistance, conducted the finishing stage (*13*). Human editing of the computer-generated alignments reduced ambiguities to about 1 in 200 nt and the autoradiograph lanes where data required proofreading were identified. Deferral of proofreading until after initial assembly saved time and reduced costs. In regions where data remained ambiguous, the finishing team requested additional data, which could involve special treatments, from the data production team. Next, a computer-aided examination for ORF's, codon usage frequencies, and similarities to database entries was used to further refine the sequence. A translated frame could often be distinguished by its codon distribution or by similarity of its predicted amino acid sequence to a known protein. The sequence was scrutinized for potential insertion or

electrophoresis through large [illegible] was used to resolve sequence, and autoradiograph films were scanned photoelectrically into computers where individual sequences were merged into overlapping contiguous segments (the assembly process). The production stage was effected by a

Transcription units were suggested by the arrangement of genes. To locate promoters of transcriptional units, a matrix search derived from in vitro measurements of Moyle *et al.* (*15*) was used to obtain consensus sequence matches which were