DE-FG02-87ER60565 (P.I. George Church, Harvard Medical School, Boston, MA)
DOE- Department of Energy HGP

Genomic Sequence Comparisons

7/1/87 – 11/14/03

Description: This project was to develop new DNA sequencing and RNA and protein quantitation methods and related genome annotation tools.

The project began in 1987 with the development of multiplex sequencing (published in Science in 1988), and one of the first automated sequencing methods. This lead to the first commercial genome sequence in 1994 and to the establishment of the main commercial participants (GTC then Agencourt) in the public DOE/NIH genome project. In collaboration with GTC we contributed to one of the first complete DOE genome sequences, in 1997, that of *Methanobacterium thermoautotropicum*), a species of great relevance to energy-rich gas production.

With key publications in 1990 & 1997, we helped pioneer the nascent field of proteomics (the term was coined in 1994) with multidimensional separations, Edman N-terminal sequencing and later mass-spectrometry. This work is ongoing with recent huge improvements in sensitivity and productivity. Augmenting the proteomics we developed some of the first approaches for using microarrays for RNA quantitation and DNA motif discovery. This has evolved into methods for integrating a wide-variety of functional genomics data including the concerted action of multiple motifs. These avalanches of genomically inspired hypotheses motivate methods for enhanced testing via semi-synthetic genomes. We pioneer this approach via array-based synthesis of oligonucleotides in 1992 and steps toward automated homologous recombination in 1997. Our recombination system has been shipped to over 1000 laboratories. These activities have all continued to improve by powers of ten to the present day. For example, our proteomics has improved from 2 peptides per day (a record in its day) to 10,000 per day; our DNA syntheses from 96 per day to 380,000; our DNA sequencing reactions from 24 per 1000 sq.cm to billions in the same space (see below)

This grant culminated in 2003 with the development of a new sequencing method, called "polymerase colony fluorescent in situ sequencing" (or polony FISSeq for short). It shares significant features reminiscent of the original multiplexing developed at the start of this grant 16 years earlier. These features include the simultaneous processing of reactions in large pools, the cycling of arrays of immobilized DNA molecules through multiple enzyme-linked steps and washes, and the precise alignment of information-rich images for each cycle. Polonies are being used for precise analysis of human genome haplotyping, RNA splicing, RNA quantiation, and human & microbial DNA sequencing.

This DOE-HGP project transitioned smoothly into a DOE GTL Center focusing on Proteomics, RNA regulation, ecological communities, and computational modeling all in the context of ocean energy metabolism and key genera including *Prochlorococcus*, *Geobacter*, & *Pseudomonas.*

Publications resulting from this grant (emphasizing the final two years)

Zhu,J, Shendure,J, Mitra, RD, Church, GM (2003) Single Molecule Profiling of Alternative Pre-mRNA Splicing. Science. 2003 Aug 8;301(5634):836-8.

Vitkup,D, Sander,C, Church,GM (2003) The Amino-acid Mutational Spectrum of Human Genetic Disease. Genome Biol. 4: R72.

King, OD, Lee, JC, Dudley, AM, Janse, DM, Church, GM, Roth, FP (2003) Predicting Phenotype from Patterns of Annotation. ISMB 2003; Bioinformatics. 2003 Jul;19 Suppl 1:I183-I189.

Grad Y, Kim J, Aach J, Hayes G, Reinhart B, Church GM, Ruvkun G. (2003) Computational and Experimental Identification of C. elegans microRNAs Molecular Cell May;11(5):1253-63.

Merritt, J, DiTonno, JR, Mitra, RD, Church, GM, Edwards, JS (2003) Functional characterization of mutant yeast PGK1 within the context of the whole cell. Nucleic Acids Research 2003 Aug 1;31(15):e84. .

Mitra,RD, Shendure,J, Olejnik,J, Olejnik,EK, and Church,GM (2003) Fluorescent in situ Sequencing on Polymerase Colonies. Analyt. Biochem. 320:55-65 Protocol & software supplements.

Mitra, RD, Butty, V, Shendure, J, Williams, BR, Housman, DE, and Church, GM (2003) Digital Genotyping and Haplotyping with Polymerase Colonies. Proc Natl Acad Sci USA. May 13;100(10):5926-31.

Selinger, DW, Saxena, RM, Cheung, KJ, Church, GM, and Rosenow, C (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation . Genome Research Feb;13(2):216-23.

Segre, D, Vitkup, D, and Church, GM (2002) Analysis of optimality in natural and perturbed metabolic networks . Proc. Nat. Acad. Sci USA 99: 15112-7. Supplement

Sudarsanam,P., Pilpel,Y, and Church, G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *S. cerevisiae* . Genome Research 12: 1723-1731

Wright, M and Church,GM (2002) An Open-source Oligonucleotide Microarray Probe Standard for Human and Mouse . Nat Biotechnol. 2002 Nov;20(11):1082-3.. Supplement

Shendure, J & Church, GM (2002) Computational discovery of sense-antisense transcription in the mouse and human genomes . Genome Biology 3:1-14 . Supplements: antisense &HumMus.

Halfon, MS, Grad, Y, Church, GM, and Michelson, AM (2002) Computation-based discovery of related transcriptional regulatory modules and motifs from a combinatorial model. Genome Research 12: 1019-1028.

Dudley, AM, Aach, J, Steffen, MA, and Church, GM (2002) Measuring absolute expression with microarrays using a calibrated reference sample and an extended signal intensity range. Proc. Nat. Acad. Sci. USA 99:7554-7559. Supplement.

Zhu, Z, Pilpel, Y, and Church, GM (2002) Computational Identification of Transcription Factor Binding Sites via a Transcription-factor-centric Clustering (TFCC) Algorithm. J. Molec. Biol. 318: 71-81

Bulyk ML, Johnson PLF, Church GM. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research 30:1255-1261.

Weber, G, Jay Shendure, J, Tanenbaum, DM, Church GM, and Meyerson M (2002) Microbial sequence identification by computational subtraction of the human transcriptome. Nature Genetics 30(2):141-2.

Smith, D.R., et al. (1997) The Complete Genome Sequence of Methanobacterium thermoautotrophicum Strain delta-H: Functional Analysis and Comparative Genomics . J. Bacteriol. Nov. 179: 7135-7155.

Link, A.J., Phillips, D. and Church, G.M. (1997) Methods for generating precise deletions and insertions in the genome of wild-type Escherichia coli: Application to open reading frame characterization. J. Bacteriol. **179:** 6228-6237.

Link, A.J., Robison, K. and Church, G.M. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli. Electrophoresis **18** (8):1259-1313 .

Tempst, P., **A. J. Link,** Riviere, L.R., Fleming, M., and Elicone, C. (1990). Internal sequence analysis of proteins separated on polyacrylamide gels at the sub-microgram level: Improved methods, applications and gene cloning strategies. Electrophoresis 11: 537-553.

Church, G.M. and Kieffer-Higgins, S. (1992) Patent WO 92/21079A1 Parallel Sequential Reactor.

Church, G.M., and Kieffer-Higgins, S. (1988) Multiplex DNA sequencing . Science **240:** 185-188.