

EXHIBIT 21

Form Approved Through 05/2004

PI: CHEE, MARK

Council: 08/2002

1 U01 HG002753-01

Dual:

IRG: ZHG1 SRC(99)

Received: 05/29/2002

Department of Health and Human Services
Public Health Services
Application
Do not exceed 56-character length restrictions, including spaces.

1. TITLE OF PROJECT Highly Parallel SNP Genotyping for a Haplotype Map					
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES (If "Yes," state number and title) Number: HG-02-005 Title: Large-Scale Genotyping for the Haplotype Map of the Human Genome					
3. PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR			New Investigator <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		
3a. NAME (Last, first, middle) Chee, Mark Stephen		3b. DEGREE(S) PhD			
3c. POSITION TITLE Vice President, Genomics		3d. MAILING ADDRESS (Street, city, state, zip code) 9885 Towne Centre Drive San Diego CA 92121-1975			
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT		E-MAIL ADDRESS: mchee@illumina.com			
3f. MAJOR SUBDIVISION Molecular Biology					
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: (858) 202-4503 FAX: (858) 202-4680					
4. HUMAN SUBJECTS RESEARCH <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes		4a. Research Exempt <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes If "Yes," Exemption No. <u>4</u>		5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
		4b. Human Subjects Assurance No.	4c. NIH-defined Phase III Clinical Trial <input type="checkbox"/> No <input type="checkbox"/> Yes	5a. If "Yes," IACUC approval Date	5b. Animal welfare assurance no
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 9/20/2002 Through 9/19/2004		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) \$7,072,747		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 7b. Total Costs (\$) \$7,310,685 8a. Direct Costs (\$) \$14,879,437 8b. Total Costs (\$) \$15,355,660	
9. APPLICANT ORGANIZATION Name Illumina, Inc. Address 9885 Towne Centre Drive San Diego, CA 92121-1975 Institutional Profile File Number (if known) 3970401			10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input type="checkbox"/> State <input type="checkbox"/> Local Private: → <input type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input checked="" type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged		
			11. ENTITY IDENTIFICATION NUMBER 33-0804655 DUNS NO. (if available) 03-330-5264 Congressional District 51st		
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name [Proprietary Info] Title [Proprietary Info] Address [Proprietary Info] Tel [Proprietary Info] FAX [Proprietary Info] E-Mail [Proprietary Info]			13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name [Proprietary Info] Title [Proprietary Info] Address [Proprietary Info] Tel [Proprietary Info] FAX [Proprietary Info] E-Mail [Proprietary Info]		
14. PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR ASSURANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. I agree to accept responsibility for the scientific conduct of the project and to provide the required progress reports if a grant is awarded as a result of this application.			SIGNATURE OF PI/PPD NAMED IN 3a. (In ink. "Per" signature not acceptable.) Requester Excluded		DATE 5/28/02
15. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.			SIGNATURE OF OFFICIAL NAMED IN 13. (In ink. "Per" signature not acceptable.) Requester Excluded		DATE 5/28/02

Principal Investigator/Program Director (Last, first, middle): Chee, Mark S.

DESCRIPTION: State the application's broad, long-term objectives and specific aims, making reference to the health relatedness of the project. Describe concisely the research design and methods for achieving these goals. Avoid summaries of past accomplishments and the use of the first person. This abstract is meant to serve as a succinct and accurate description of the proposed work when separated from the application. If the application is funded, this description, as is, will become public information. Therefore, do not include proprietary/confidential information. **DO NOT EXCEED THE SPACE PROVIDED.**

The majority of genes involved in common disease remain unknown. Discovery of these genes will transform our knowledge of the genetic contribution to human disease, and lead to the provision of new genetic screens, and underpin research into new cures or improved lifestyles. A leading strategy for their discovery is to test specific sequence variants for association with a measurable phenotype, and from this to identify the causative variant and hence the gene involved. An important first step is to create a haplotype map of the genome and identify a minimal set of SNPs that can be used to detect common haplotype patterns in multiple populations. This will enable comprehensive genome-wide genetic association studies, potentially revolutionizing the search for the genetic basis of common diseases. The overall aim of this project is to select common variants in the form of single nucleotide polymorphisms (SNP) from the human genome sequence, and to carry out large-scale genotyping with the goal of creating such a haplotype map for a significant fraction of the human genome. Genotyping will be carried out using a novel, parallel large-scale genotyping system that combines a highly multiplexed assay format, a miniaturized bead-based array platform, and positively-tracked, LIMS-based, modular automation. The system has a base capacity of ~ 1,000,000 genotypes per day, and is easily scaled up to much higher capacities. It will be used to develop and screen assays for 400,000 SNPs. The SNPs will be genotyped in a set of samples representing African, Asian, and Caucasian populations, and will provide a data set of ~ 74 million genotypes for analysis. The data will be used to define haplotype patterns that are common in each population, and to identify a specific set of SNPs ("tag SNPs") which will be maximally informative for future genome wide association studies to investigate the role of common variants in common disease. This study will form part of an international collaborative programme (the "HapMap" project) which will make all the information relating to this work freely available in the public domain.

PERFORMANCE SITE(S) (organization, city, state)

Illumina, Inc., San Diego, CA

KEY PERSONNEL. See instructions. Use continuation pages as needed to provide the required information in the format shown below. Start with Principal Investigator. List all other key personnel in alphabetical order, last name first.

Name	Organization	Role on Project
Mark Chee	Illumina, Inc.	PI

Proprietary Info

Disclosure Permission Statement. Applicable to SBIR/STTR Only. See instructions. Yes No

Principal Investigator/Program Director (Last, first, middle): Chee, Mark S.

The name of the principal investigator/program director must be provided at the top of each printed page and each continuation page.

**RESEARCH GRANT
TABLE OF CONTENTS**

	<i>Page Numbers</i>
Face Page	1
Description, Performance Sites, and Personnel	2-
Table of Contents	3
Detailed Budget for Initial Budget Period (or Modular Budget).....	4
Budget for Entire Proposed Period of Support (not applicable with Modular Budget).....	5-
Budgets Pertaining to Consortium/Contractual Arrangements (not applicable with Modular Budget)	
Biographical Sketch—Principal Investigator/Program Director (<i>Not to exceed four pages</i>)	7-
Other Biographical Sketches (<i>Not to exceed four pages for each – See instructions</i>)	11-
Resources	19
 Research Plan	
Introduction to Revised Application (<i>Not to exceed 3 pages</i>).....	
Introduction to Supplemental Application (<i>Not to exceed one page</i>).....	
A. Specific Aims	20
B. Background and Significance.....	21
C. Preliminary Studies/Progress Report/ Phase I Progress Report (SBIR/STTR Phase II ONLY)	22
D. Research Design and Methods.....	32
E. Human Subjects	43
Protection of Human Subjects (Required if Item 4 on the Face Page is marked "Yes")	43
Inclusion of Women (Required if Item 4 on the Face Page is marked "Yes")	43
Inclusion of Minorities (Required if Item 4 on the Face Page is marked "Yes")	44
Inclusion of Children (Required if Item 4 on the Face Page is marked "Yes")	
Data and Safety Monitoring Plan (Required if Item 4 on the Face Page is marked "Yes" <u>and</u> a Phase I, II, or III clinical trial is proposed).....	
F. Vertebrate Animals	44
G. Literature Cited	44
H. Consortium/Contractual Arrangements	
I. Letters of Support (e.g., Consultants).....	47
J. Product Development Plan (SBIR/STTR Phase II and Fast-Track ONLY)	
 Checklist	 50

(Items A-D: not to exceed 25 pages*)
* SBIR/STTR Phase I: Items A-D limited to 15 pages.

Appendix (*Five collated sets. No page numbering necessary for Appendix.*)
 Appendices NOT PERMITTED for Phase I SBIR/STTR unless specifically solicited. X Check if Appendix is Included

Number of publications and manuscripts accepted for publication (*not to exceed 10*) _____
 Other items (list): _____

Principal Investigator/Program Director (Last, first, middle): Chee, Mark S.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY					FROM 9/20/2002	THROUGH 9/19/2003	
PERSONNEL <i>(Applicant organization only)</i>			TYPE APPT. <i>(month)</i>	% EFFORT ON PROJ. <i>% Effort</i>	INST. BASE SALARY	DOLLAR AMOUNT REQUESTED <i>(omit cents)</i>	
NAME	ROLE ON PROJECT	EFFORT				SALARY REQUESTED	FRINGE BENEFITS
Mark Chee	Principal Investigator						
Proprietary Info			Proprietary Info	Institutional Base Salary	Proprietary Info		
SUBTOTALS →							
CONSULTANT COSTS							
Proprietary Info							0
EQUIPMENT <i>(itemize)</i>							
Proprietary Info							Proprietary Info
SUPPLIES <i>(itemize by category)</i>							
Proprietary Info							Proprietary Info
TRAVEL							
Travel to attend HapMap Network meetings (2 individuals x 2 meetings / yr)							6,000
PATIENT CARE COSTS							
INPATIENT							
OUTPATIENT							
ALTERATIONS AND RENOVATIONS <i>(itemize by category)</i>							
OTHER EXPENSES <i>(itemize by category)</i>							
Proprietary Info							Proprietary Info
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						\$7,072,747	
CONSORTIUM/CONTRACTUAL COSTS		DIRECT COSTS					
		FACILITIES AND ADMINISTRATIVE COSTS					
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD <i>(Item 7a, Face Page)</i> →						\$7,072,747	

Principal Investigator/Program Director (Last, first, middle): Chee, Mark S.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS	INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED				
		2nd	3rd	4th	5th	
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>	Proprietary Info					
CONSULTANT COSTS						
EQUIPMENT						
SUPPLIES						
TRAVEL						
PATIENT CARE COSTS		INPATIENT				
		OUTPATIENT				
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES						
SUBTOTAL DIRECT COSTS						
CONSORTIUM/ CONTRACTUAL COSTS		DIRECT				
		F&A				
TOTAL DIRECT COSTS		7,072,747	7,806,690			

TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD (Item 8a, Face Page) _____ **\$ 14,879,437**

**SBIR/STTR Only
Fee Requested**

SBIR/STTR Only: Total Fee Requested for Entire Proposed Project Period
(Add Total Fee amount to "Total direct costs for entire proposed project period" above and Total F&A/indirect costs from Checklist Form Page, and enter these as "Costs Requested for Proposed Period of Support on Face Page, Item 8b.)

\$

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Personnel

Mark Chee, Ph.D., % Effort Principal Investigator. Responsible for the overall direction, supervision and coordination of the project, and for carrying out the PI's responsibilities as defined in the HapMap RFA.

Proprietary Info

Principal Investigator/Program Director (Last, first, middle): Chee, Mark S.

Proprietary Info

Consultants

Proprietary Info

Equipment

Proprietary Info

Supplies

Arrays, reagents, disposable lab supplies for one scientist to carry out cost reduction experiments.

Travel

For the PI and an additional individual to attend two HapMap Network meetings per year.

Other Expenses

Proprietary Info

YEAR 2

Personnel

Proprietary Info

Other Expenses

The amount of assay development and genotyping is increased somewhat in Year 2.

Principal Investigator: Chee, Mark S.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed for Form Page 2.
Follow the sample format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME		POSITION TITLE		
Chee, Mark Stepen		Vice President, Genomics		
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)				
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY	
University of New South Wales, Australia	BSc. Hons	1985	Biochemistry	
University of Cambridge, United Kingdom	Ph.D.	1991	Molecular Biology	

A. POSITIONS AND HONORS**Professional Experience and Employment**

1990-1991	Postdoctoral Fellow	Molecular Biology	MRC Laboratory of Molecular Biology
1992	Postdoctoral Fellow	Molecular Biology	Stanford University and Affymax Research Institute
1993-1995	Staff Scientist	Molecular Biology	Affymetrix, Inc.
1996	Senior Scientist	Molecular Genetics	Affymetrix, Inc.
4/97-7/97	Director	Genetics Research	Affymetrix, Inc.
6/98-present	Vice President	Genomics	Illumina, Inc.

Professional Memberships

American Association for the Advancement of Science
American Society of Human Genetics
Society for General Microbiology (U.K.)

Awards

Commonwealth Scholarship at University of Cambridge, 1986-1989
Applied Biosystems Fellowship at MRC Laboratory of Molecular Biology, 1989-1991

B. SELECTED PEER-REVIEWED PUBLICATIONS

(Selected from 25 peer-reviewed publications)

- 1) Chee, M. S., Rudolph, S. A., Plachter, B., Barrell, B. G. & Jahn, G. (1989) Identification of the major capsid protein gene of human cytomegalovirus. *Journal of Virology* **63**, 1345-1353.
- 2) Chee, M. S., Lawrence, G. L. & Barrell, B. G. (1989) Alpha-, beta- and gammaherpesviruses encode a putative phosphotransferase. *Journal of General Virology* **70**, 1151-60.
- 3) Lawrence, G. L., Chee, M., Craxton, M. A., Gompels, U. A., Honess, R. W. & Barrell, B. G. (1990) Human herpesvirus-6 is closely related to human cytomegalovirus. *Journal of Virology* **64**, 287-299.
- 4) Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Horsnell, T., Hutchison III, C. A., Kouzarides, T., Martignetti, J. A., Satchwell, S. C., Tomlinson, P., Weston, K. M. & Barrell, B. G. (1990) Analysis of the protein coding content of the sequence of human cytomegalovirus strain AD169. *Current Topics in Microbiology & Immunology* **154**, 125-169.
- 5) Chee, M. S., Satchwell, S. C., Preddie, E., Weston, K. M. & Barrell, B. G. (1990) Human cytomegalovirus encodes three G protein-coupled receptor homologues. *Nature* **344**, 774-777.
- 6) Littler, E., Stuart, A. D. & Chee, M. S. (1992) Human cytomegalovirus UL97 open reading frame encodes a protein that phosphorylates the antiviral nucleoside analogue ganciclovir. *Nature* **358**, 160-162.

Principal Investigator: Chee, Mark S.

- 7) Smith, V., Craxton, M., Bankier, A. T., Brown, C. M., Rawlinson, W. D., Chee, M. S. & Barrell, B. G. (1993) Microtitre methods for the preparation and fluorescent sequencing of M13 clones. *Methods in Enzymology*, **218**, 173-187.
- 8) Kozal, M. J., Shah, N., Shen, N., Yang, R., Fucini, R., Merigan, T. C., Richman, D. D., Morris, M. S., Hubbell, E., Chee, M. S. & Gingeras, T. R. (1996) Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nature Medicine* **2**, 753-759.
- 9) Chee, M. S., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. A. (1996) Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-614.
- 10) Hacia, J. G., Brody, L. C., Chee, M. S., Fodor, S. P. A. & Collins, F. S. (1996) Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-color fluorescence analysis. *Nature Genetics* **14**, 441-447.
- 11) Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E. L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675-1680.
- 12) Mackett, M., Stewart, J. P., de V Pepper, S., Chee, M., Efstathiou, S., Nash, A. A., Arrand, J. R. (1997) Genetic content and preliminary transcriptional analysis of a representative region of murine gammaherpesvirus 68. *Journal of General Virology* **78**, 1425-33.
- 13) Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, MacDonald S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M. & Lander, E. S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082.
- 14) Gunderson, K., Huang, X. C., Morris, M. S., Lipshutz, R. J., Lockhart, D. J. & Chee, M. S. (1998) Mutation identification by hybridization to complete n-mer DNA arrays. *Genome Research* **8**, 1142-1153.
- 15) Gentalen, E. & Chee, M. S. (1999) A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays. *Nucleic Acids Research* **27**, 1485-1491.
- 16) Mei, R., Galipeau, P. C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R. K., Chee, M. S., Reid, B. J. & Lockhart D. J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Research* **8**, 1126-1137.
- 17) Yeakley, J. M., Fan, J.-B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S. & Fu, X.-D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology* **20**, 353-358.

US Patents

- 1) US Patent # **5,837,832**. Arrays of nucleic acid probes on biological chips. Chee, M. S., Cronin, M. T., Fodor, S. P. A., Huang, X. C., Hubbell, E. A., Lipshutz, R. J., Lobban, P. E., Morris, M. S. & Sheldon, E. L.
- 2) US Patent # **5,856,104**. Polymorphisms in the glucose-6-phosphate dehydrogenase locus. Chee, M. S. & Fan, J.-B.
- 3) US Patent # **5,974,164**. Computer-aided visualization and analysis system for sequence evaluation. Chee, M. S.
- 4) US Patent # **6,040,138**. Expression monitoring by hybridization to high density oligonucleotide arrays. Lockhart, D. J., Brown, E. L., Wong, G. G., Chee, M. S. & Gingeras, T. R.
- 5) US Patent # **5,861,242**. Array of nucleic acid probes on biological chips for diagnosis of HIV and methods of using the same. Chee, M. S., Gingeras, T. R., Fodor, S. P. A., Hubbell, E. A. & Morris, M. S.
- 6) US Patent # **6,013,440**. Nucleic acid affinity columns. Lipshutz, R. J., Morris, M. S., Chee, M. S. & Gingeras, T. R.
- 7) US Patent # **6,027,880**. Arrays of nucleic acid probes and methods of using the same for detecting cystic fibrosis. Cronin, M. T., Miyada, C. G., Hubbell, E. A., Chee, M. S., Fodor, S. P. A., Huang, X. C., Lipshutz, R. J., Lobban, P. E., Morris, M. S. & Sheldon, E. L.

Principal Investigator: Chee, Mark S.

- 8) US Patent # **6,050,719**. Rotational mixing method using a cartridge having a narrow interior. Winkler, J. L., **Chee, M. S.** & Lockhart, D. J.
- 9) US Patent # **6,156,501**. Arrays of modified nucleic acid probes and methods of use. McGall, G. H., Miyada, C. G., Cronin, M. T., Tan, J. D. & **Chee, M. S.**
- 10) US Patent # **6,228,575**. Chip-based species identification and phenotypic characterization of microorganisms. Gingeras, T. R., Mack, D., **Chee, M. S.**, Berno, A. J., Stryer, L., Ghandour, G. & Wang, C.
- 11) US Patent # **6,238,862**. Methods for testing oligonucleotide arrays. McGall, G. H., Barone, A. D., Diggelmann, M., Lockhart, D. J., Caviani Pease, A. M. & **Chee, M. S.**
- 12) US Patent # **6,242,180**. Computer-aided visualization and analysis system for sequence evaluation. **Chee, M. S.**
- 13) US Patent # **6,280,950**. Nucleic acid affinity columns. Lipshutz, R. J., Morris, M. S., **Chee, M. S.**, Gingeras, T. R.
- 14) US Patent # **6,306,643**. Methods of using an array of pooled probes in genetic analysis. Gentalen, E. & **Chee, M. S.**
- 15) US Patent # **6,309,823**. Arrays of nucleic acid probes for analyzing biotransformation genes and methods of using the same. Cronin, M. T., Miyada, C. G., Hubbell, E. A., **Chee, M. S.**, Fodor, S. P. A., Huang, X. C., Lipshutz, R. J., Lobban, P. E., Morris, M. S. & Sheldon, E. L.
- 16) US Patent # **6,342,355**. Probe-based analysis of heterozygous mutations using two-color labeling. Hacia, J. G., **Chee, M. S.** & Collins, F. S.
- 17) US Patent # **6,344,316**. Nucleic acid analysis techniques. Lockhart, D. J., **Chee, M. S.**, Gunderson, K., Lai, C., Wodicka, L., Cronin, M. T., Lee, D., Tran, H. M. & Matsuzaki, H.
- 18) US Patent # **6,355,431**. Detection of nucleic acid amplification reactions using bead arrays. **Chee, M. S.** & Gunderson, K.
- 19) US Patent # **6,368,799**. Method to detect polymorphisms and monitor allelic expression employing a probe array. **Chee, M. S.**

C. RESEARCH SUPPORT

Ongoing Research Support

5R44 HG02003-01 Chee (PI)

9/25/2000 -> 8/31/2002

NIH/NHGRI

Randomly Ordered DNA Arrays for SNP Genotyping

The goals of this project are to develop a 2,000 marker SNP genotyping system based on randomly ordered arrays and a microtiter plate-based system for genotyping large numbers of samples.

Role: Principal Investigator

2R44 CA81952-02 Chee (PI)

07/01/2001 -> 06/30/2003

NIH/NCI

Random Arrays for Gene Expression Profiling

The goal of this project is to develop random array technology for analysis of mRNA expression.

Role: Principal Investigator

R33 CA88197 Chee (PI)

2/1/2002 -> 1/31/2004

NIH/NCI

Protein Profiling Arrays

The goal of this project is to develop bead array technology for simultaneously measuring many proteins and their post-translational modifications in small volumes of cells or biological fluids.

Role: Principal Investigator

R44 CA83398-02 Chee (PI)

4/1/2002 -> 3/31/2004

NIH/NCI

Principal Investigator: Chee, Mark S.

Parallel Array Processor

The goal of this project is to develop a system for parallel processing of arrays.

Role: Principal Investigator

Completed Research Support

(Last 3 Years)

1 R21 HG01911-01 Chee (PI)

02/01/1999 -> 11/30/1999

NIH/NHGRI

Decoding Randomly Ordered Arrays

The goal of this project was to demonstrate the feasibility of decoding randomly assembled arrays by sequential hybridization of pools of fluorescently labeled oligonucleotide probes.

Role: Principal Investigator

DE-FG02-00ER62923 Walt (PI)

6/01/1999 -> 2/28/2001

DOE

Time-Resolved Sequence Analysis on High Density Fiberoptic DNA Probe Arrays

The goal of this project was to develop methods of analyzing nucleic acid sequence using an array of oligonucleotide probes attached to beads.

Role: Co-Investigator

1 R43 GM62094-01 Chee (PI)

07/01/2000 -> 5/31/2001

NIH/NIGMS

Invader Arrays for Nucleic Acid Analysis

The goal of this project was to assess the feasibility of combining the InvaderTM assay for nucleic acids with a miniaturized bead array format.

Role: Principal Investigator

1 R43 GM61511-01 Lebl (PI)

7/01/2000 -> 12/31/2000

NIH/NIGMS

Automated DNA Synthesizer Using Tilted Plate Technology

The goals of this project were to design, construct and evaluate an instrument for oligonucleotide synthesis.

Role: Co-Investigator

1 R43 HG02119-01 Chee (PI)

02/01/2000 -> 10/31/2000

NIH/NIGMS

Pyrosequencing Arrays

The goal of this project was to assess the feasibility of combining the PyrosequencingTM assay for nucleic acids with a miniaturized bead array format.

Role: Principal Investigator

Principal Investigator: Chee, Mark S.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed on Form Page 2.
Photocopy this page or follow this format for each person.

Proprietary Info



Principal Investigator: Chee, Mark S.

Proprietary Info



Principal Investigator: Chee, Mark S.

BIOGRAPHICAL SKETCH

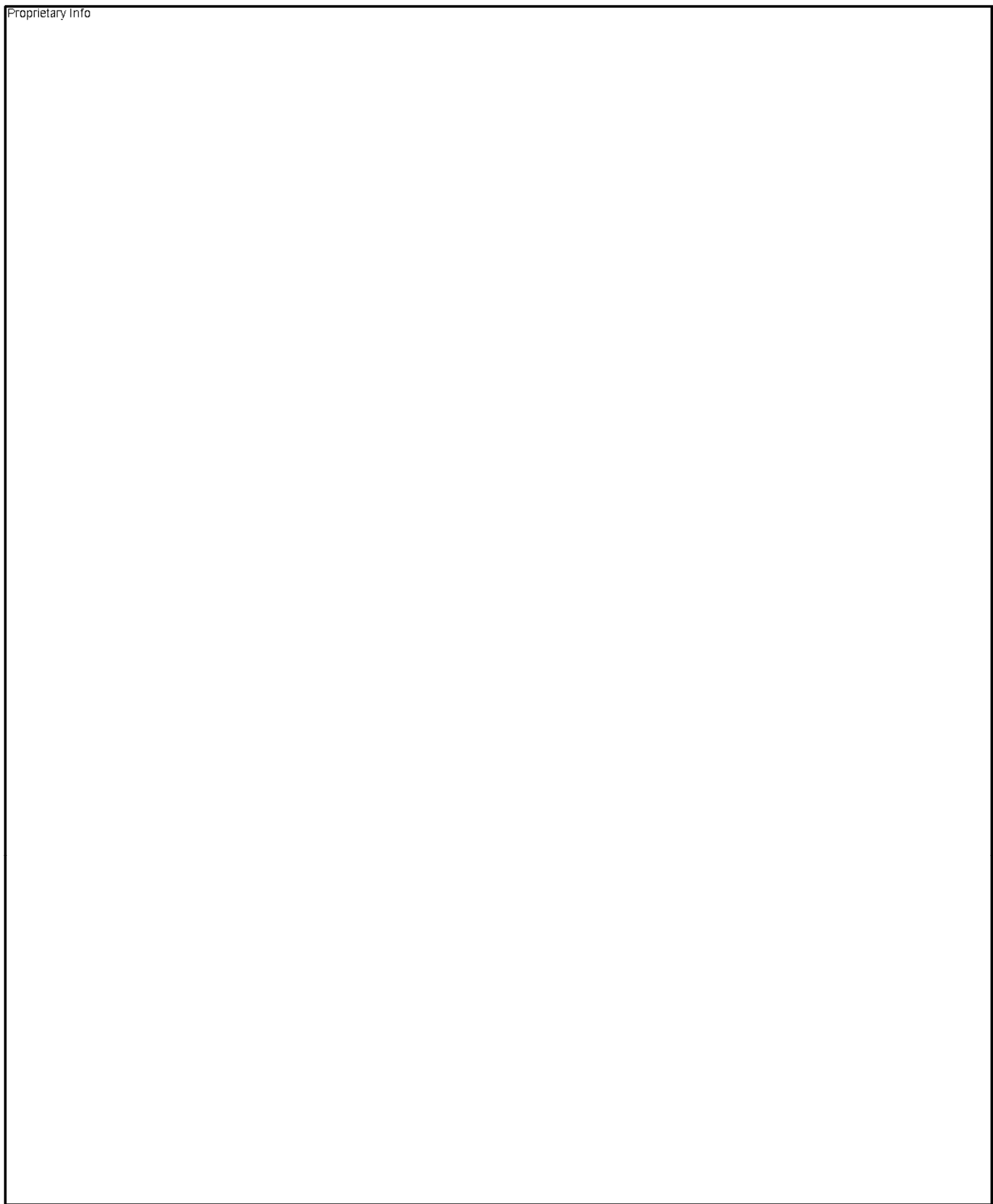
Provide the following information for the key personnel in the order listed on Form Page 2.
Photocopy this page or follow this format for each person.

Proprietary Info



Principal Investigator: Chee, Mark S.

Proprietary Info



Principal Investigator: Chee, Mark S.

Proprietary Info



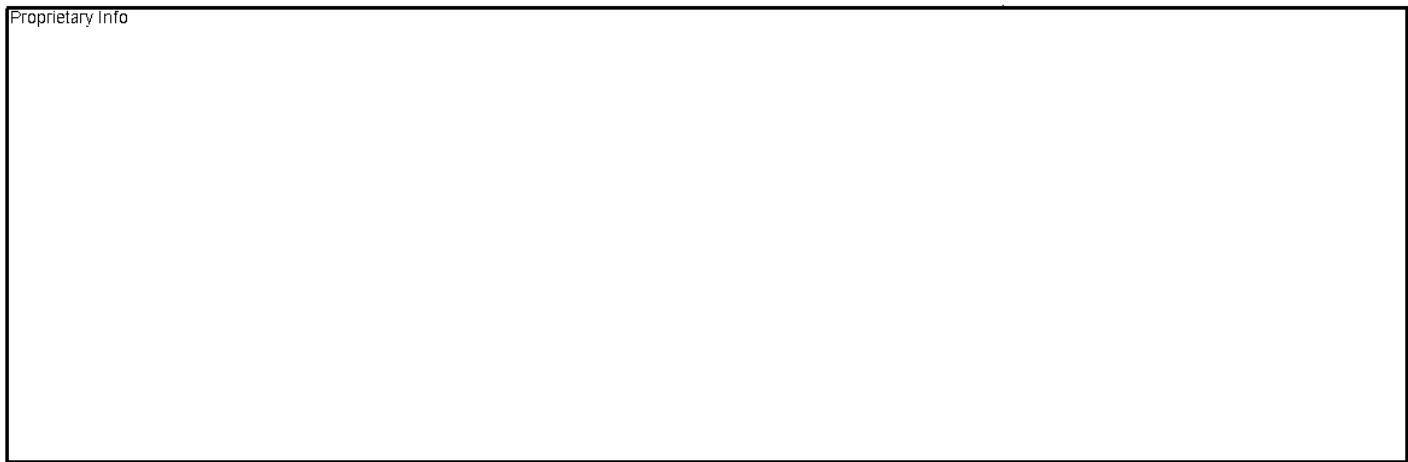
Principal Investigator/Program Director (Last, first, middle):

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed for Form Page 2.
Follow the sample format for each person. **DO NOT EXCEED FOUR PAGES.**

Proprietary Info

Proprietary Info



Principal Investigator: Chee, Mark S.

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed on Form Page 2.
Photocopy this page or follow this format for each person.

Proprietary Info



Principal Investigator/Program Director (Last, first, middle):

Proprietary Info

RESEARCH PLAN

A. SPECIFIC AIMS

The overall aim of this project is to carry out large-scale genotyping with the goal of defining common haplotype blocks for a significant fraction of the human genome.

1. Assays will be developed and screened for a total of 400,000 SNPs, to obtain an estimated Proprietary Info For maximum efficiency, we will define haplotype blocks using a hierarchical strategy involving several rounds of SNP assay development, with the optimal resolution for each round of screening to be determined prior to the start of the project. Proprietary Info

Proprietary Info

2. The 400,000 SNPs will be genotyped in Proprietary Info representing African, Asian, and Caucasian populations, Proprietary Info We anticipate doing this at levels of Proprietary Info throughout the entire process. A quality score will be provided for each genotype, with a projected accuracy of 99.7% for successful genotypes.

3. Since the actual yield of informative SNPs depends on several variables, an initial goal of the project is to optimize our assay development process based on empirical data from a pilot set of 40,000 SNPs. Proprietary Info

Proprietary Info

Proprietary Info

4. A major advantage of our highly multiplexed genotyping technology is that additional samples can be genotyped very cost effectively. Proprietary Info

Proprietary Info

Proprietary Info

5. Proprietary Info

This proposal requests funding for genotyping and analysis to be carried out at Illumina. Proprietary Info

Proprietary Info

Proprietary Info

Genotyping will be carried out using a novel, parallel large-scale genotyping system that

combines a highly multiplexed assay format; a miniaturized bead-based array platform; and positively-tracked, LIMS-based, modular automation. The system has a base capacity of ~ 1,000,000 genotypes per day, and is easily scaled up to much higher capacities.

B. BACKGROUND AND SIGNIFICANCE

B.1 Rationale for Development of a Haplotype Map

It has been widely proposed that the genetic component of important medical phenotypes such as risk of common disease and drug response may be identified by studies to determine the association of specific sequence variants such as single nucleotide polymorphism (SNP) alleles with a specific phenotype (association studies). While candidate genes offer an early start-point, they require prior knowledge of gene function to enable the hypothesis to be made of involvement of the candidate gene in the disease. The majority of genes involved in common disease remain unknown. It is envisaged that genome-wide testing through association studies will be a leading strategy for their discovery. Availability of common SNPs throughout the genome provides the basis for such association studies, but the ability to detect association between an allele of a SNP in the genome map, and a causative functional variant (FV), relies on linkage disequilibrium between the SNP and the causative variant. Early studies indicate the existence of variable patterns of LD in the genome, averaging 50kb and including some extensive regions (hundreds of kb) (Daly et al. 2001; Dawson; Gabriel et al. 2002; Jeffreys et al. 2001; Patil et al. 2001; Reich et al. 2001). The variable nature of LD further indicates the need to establish a genome map of LD and common haplotypes empirically. This may be done by large scale genotyping of SNPs in multiple DNA samples chosen to be representative of the human population, followed by measurement of inter-marker LD and definition of the identity and frequency of all haplotype patterns, or blocks. Furthermore, knowledge of such haplotype patterns can be used to select a minimal subset of SNPs that can be used to detect all the previously determined haplotype patterns in subsequent population studies. These SNPs have been referred to as "haplotype tag" or htSNPs.

Studies to date indicate that there are a limited number of common patterns in blocks, and that there is significant conservation of block structure across different populations, presumably reflecting non-random ancestral recombination events which disrupt LD in limited regions of the genome (i.e. between blocks). The majority of the genome (up to 95%), by contrast, may have blocks of high LD (little or no ancestral recombination) and this structure can be defined using a sufficiently high density of SNPs.

B.2 The Role of Technology in Creating a Haplotype Map and Enabling Large-Scale Genetic Studies

The scale of SNP genotyping needed for creating a haplotype map, and using it widely in subsequent genetic studies, is orders of magnitude greater than has been required for conventional family-based linkage mapping. Realizing the vision of comprehensive genome-wide genetic association studies will require a SNP genotyping system that combines very high throughput and accuracy with very low cost per SNP analysis. This combination has so far proved to be elusive (Kwok 2001; Syvanen 2001) Our approach to this problem has been to combine a miniaturized assay platform, a high level of multiplexing, and flexible automation.

In Section C, we will show data that demonstrate a highly parallel genotyping technology that already can operate at a throughput of hundreds of thousands to millions of genotypes per day, with only

modest requirements in capital equipment and labor. Furthermore, throughput can easily be scaled to much higher levels. Importantly, this technology is capable of delivering a commercial price point of a penny per genotype, assuming sufficiently large market demand to justify pricing at that level.

By participating in the development of the haplotype map, we aim to use the BeadArray technology to accelerate the completion of the project, and also to provide the means to improve the resolution and utility of the finished map. By minimizing the cost and maximizing the rate of genotyping, it will be possible to reach the currently defined endpoint of the project more rapidly and at lower cost. As a result, it will be feasible to analyze more SNPs and also more DNA samples, as indicated by the results of further pilot studies. Secondly, since the creation of the map demands assay development on a large scale, the project would help optimize this part of our process. This will allow us in the future to provide on a commercial basis genotyping systems that are optimized for low cost, high-throughput genotyping, and also for assay development. Finally, our current focus has been on genotyping technology, an area where we have been able to concentrate considerable expertise representing a range of scientific and engineering disciplines, including informatics and software development. By participating in the haplotype mapping project, we hope to share our expertise in these areas with those of our collaborators and others in the project who have complementary expertise in population genetics and application of the technology to disease studies. This would enable us both to learn more about the specific problems posed by this area of research, and to contribute our informatics expertise to the development of algorithms and software to help enable large-scale genetic studies of common variants and common disease that are truly cost-effective and comprehensive.

C. PRELIMINARY STUDIES

C.1 Experience in Technology Development

In a period of ~ 4 years, we have developed, among other technologies:

- a) A novel array technology platform, including setting up a state-of-the-art manufacturing facility with the capacity to manufacture many thousands of 96-array matrices per year,
- b) A high-throughput genotyping technology based on the BeadArray™ platform,
- c) An integrated, high-capacity genotyping system that is now in commercial use, with a capacity of ~ 1,000,000 genotypes per day,
- d) A high-throughput oligonucleotide manufacturing technology and production facility, providing oligos commercially at \$0.16 per base for 25 nmol scale synthesis (and at \$0.11 per base for this project), and
- e) High-performance array imaging systems for internal and commercial use.

In the process, Illumina has grown from 3 to over 200 employees, with expertise in assay development and molecular biology, software development, data analysis and bioinformatics, chemistry, engineering, automation, manufacturing process development and manufacturing, and production genotyping, as well as infrastructure for technology commercialization.

In Section C we provide an overview of our genotyping system and its capabilities, and summarize our genotyping experience using this system in a production environment. Proprietary Info

Proprietary Info

Proprietary Info Section D describes the assay and array technologies of the system in more detail.

C.1.1 Integrated Genotyping System

The philosophy for the creation of our production system is modularity (Fig. 1). We minimize the amount of custom equipment that is required by purchasing off the shelf robotics, ovens and thermal cyclers. We use a laboratory information management system (LIMS) to achieve informatic integration of all steps of the genotyping process. The LIMS allows us to check and control the sample flow through the production system and provides flexibility for quick redesign and process development. Modularity in the design of the production process allows us to increase capacity very quickly. We would need only to purchase more robots, ovens and thermocyclers to increase capacity, relying on LIMS to tie the production process together.

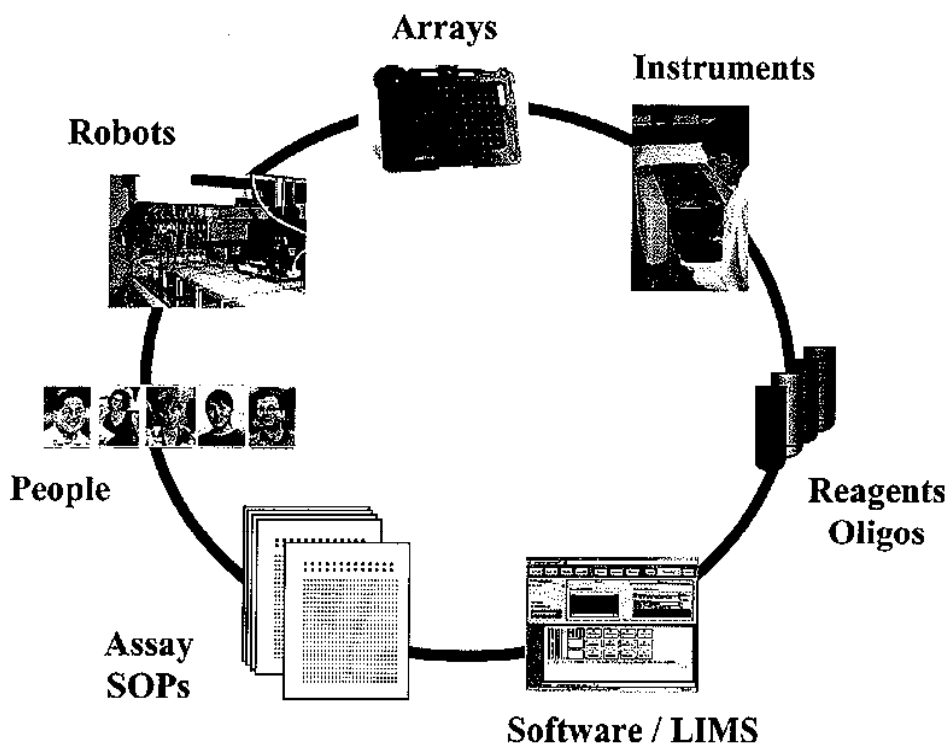


Figure 1. Informatically integrated high-throughput genotyping system based on BeadArray technology. The system is modular, and integrated using barcode reading and LIMS supervision. Miniaturization is achieved by using a fiber-optic bundle as a substrate for the highest-density microarray available today. Ninety-six of these arrays are held together in a matrix (Array of Arrays™ matrix) that matches the spacing of a standard 96-well microplate. In order to use efficiently the high capacity of the platform, the SNP assays developed at Illumina have been designed for a high level of multiplexing. Currently, Illumina's production genotyping system routinely multiplexes 288 SNP assays in each well of a microplate, and we have recently achieved excellent results with 1,152-plex assays.

Any part of the system can be upgraded easily, since components can be added or replaced without disrupting the design. The modularity also minimizes downtime of the production system. Each robot is able to perform every process in production. If one robot were to break, another robot is able to perform the same process in production. Contrast this to a fully integrated robotic system, where serialized modules perform a single task. If any part of that integrated system were to break the whole system would be idled.

Informatic integration means that all robotic pipetting steps are controlled by LIMS in real time. When the operator requests a robotic pipetting procedure, the LIMS activates a barcode reader to identify samples and reagents on the robot bed. It knows what procedures have already been carried out on the samples, and can therefore determine if the requested procedure is appropriate. If so, the LIMS activates the robot and carries out the procedure. This allows us to process large numbers of samples accurately. In addition, the LIMS manages the actual genotype calling by interacting directly with software that calls genotypes, and ensures that genotyping results are associated with the correct samples. The LIMS is also used to monitor quality metrics and correlate these with process variables, and to automate and manage oligonucleotide design and pooling for multiplexed assay development.

C.1.2 Production Environment and Genotyping Capacity

The production environment has been in commercial operation for over 6 months, and in this time has been highly reliable and consistent. The production facility operates in a well-defined way. Trained personnel run the genotyping system described above, using standard operating procedures. Reagents are quality-controlled prior to use, and pre-PCR and post-PCR steps are carried out in separate laboratories. The pre-PCR facility is access-controlled and under positive pressure. Transfer of materials is one-way from pre-PCR to post-PCR through a portal. These precautions help to maintain a contamination-free environment.

Proprietary Info

Proprietary Info

Table 1. Throughput of genotyping system. The genotyping capacity is shown as a function of level of multiplexing (loci / array) and the number of 96-array matrices processed per day.

Most of the genotyping to date has been devoted to internal use, mainly for optimization experiments - many millions of genotypes have been generated in exploring and optimizing a range of variables for each step of the genotyping process.

Proprietary Info

Proprietary Info

C.2 Measures of Quality

Genotype quality is quantified at multiple levels in our production process. We have developed algorithms (GenTrain™ and GenCall™) that are used to both call genotypes automatically and to estimate the accuracy of our genotype calling process. Automated genotype calling is necessary for consistency and objectivity, and also because it is simply impractical to have any human involvement given the large amounts of data generated.

Proprietary Info

Proprietary Info GenTrain and GenCall are used to generate a genotype and give it a score (GenCall score). We use the GenCall score

Proprietary Info

Proprietary Info to measure performance of the production process. The utility of the GenCall score is verified using measures of reproducibility and accuracy (Section C.2.1).

We leverage the use of our laboratory information management system (LIMS) to tell us about the performance of each component of the production process.

Proprietary Info

Proprietary Info

Proprietary Info Our genotyping process uses two dyes, Cy3 and Cy5, to tell us about the genotype call for each locus. The ratios of the signal intensity for these two dyes tell us about the amount of each allele present in a sample.

Proprietary Info

C.2.1 Measuring Accuracy

The automatic computation of a GenCall score for each genotype provides an objective measure of quality that is of great benefit in a large-scale genotyping project. The usefulness of the GenCall score derives from the work we have done to correlate GenCall scores with accuracy, based on efficient ways of estimating accuracy in large data sets.

As a result of our need to analyze large data sets efficiently, some of our approaches for analysis and reporting are more advanced than those typically used for small-scale genotyping. Proprietary Info

Proprietary Info

Proprietary Info

Definitions and procedures for evaluating genotyping accuracy are as follows:

1) *Concordance* refers to comparing two sets of genotyping results obtained by different methods. Consequently, the resolution of discordant results between the two methods requires analysis by a third technology, such as DNA sequencing.

2) *Reproducibility* simply refers to reproducing genotyping calls on replicate DNA samples. Proprietary Info

Proprietary Info

3) *Strand correlation* refers to comparing genotype calls for each SNP on both DNA strands. The assumptions for this method of calculating accuracy are that, a) for any particular locus, the genotype on the top strand matches the genotype on the bottom strand, and b) the ability to develop a SNP assay for a particular locus is strand independent. We believe that the second assumption is not always true; instead, if a robust assay is developed on one strand, it is more likely that the assay will develop well on the other strand. As a result, strand correlation may slightly overestimate accuracy. Nevertheless, we have found strand correlation to be an effective way to monitor accuracy in studies to test and refine system performance.

GenCall scores have been shown to correlate with accuracy as measured by reproducibility and by strand correlation. This has been verified by an independent study of quality, using Mendelian error-checking, described in Section C.3. In comparing accuracy across platforms or studies, it is important to consider the call rate in association with accuracy. By lowering the call rate and excluding low quality results, accuracy can be increased. Because GenCall score correlates with accuracy, the user is able to choose a data acceptance threshold based on GenCall score. In this way the analyzer of the data can also balance completeness of data versus accuracy. For instance, one can lower the call rate and analyze fewer data, but have a higher assurance of accuracy. Our results to date have yielded high call rates in conjunction with high accuracy (Section C.3).

Another critical factor that affects accuracy is the quality of the DNA sample used in the assay. Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

In this study a large portion of inaccurate calls can be assigned to relatively few DNAs.

Proprietary Info

Proprietary Info

C.3 Independent Assessment of Quality

Proprietary Info

Proprietary Info

Proprietary Info At the same time it was possible to assess overall efficiencies of assay development using SNPs selected from the publicly available map constructed by the International SNP Working Group (Sachidanandam et al. 2001). 1100 SNPs were selected Proprietary Info from the map, each with at least 100 bases of unique sequence flanking the polymorphic base, and spaced at intervals of approximately 1 SNP / kb in a 10Mb region of the finished genomic sequence of human chromosome 20 (Deloukas et al. 2001). Proprietary Info

Proprietary Info Each SNP multiplex was tested on Proprietary Info

Proprietary Info (Dawson et al., Nature in press), Proprietary Info

Proprietary Info The rate of conversion of SNPs selected in the map to polymorphic assays at this minor allele frequency was identical to that obtained by a similar analysis using the Sequenom platform.

Proprietary Info Errors were checked using the following three methods. Proprietary Info

Proprietary Info The resulting duplicate genotypes were compared (i.e. a test of reproducibility, as defined above). All discrepancies were scored and correlated with Gencall scores. Proprietary Info results (see table below) show that 95.7% of the genotypes (all GenCall scores > 0.5) for which a duplicate

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info **Second** Proprietary Info
Proprietary Info **As before, all discrepancies were scored and correlated with GenCall scores.** Proprietary Info
Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info **Third,** Proprietary Info
Proprietary Info
Proprietary Info

In conclusion, the data generated from a representative set of SNPs selected from the publicly available map is of sufficiently high accuracy for use in measurements of LD and haplotype blocks. Furthermore, the use of well-established strategies for error measurement such as those described here provide accurate correlation with the GenCall score.

C.4 Multiplexing

By using objective measures of quality described above, we have optimized assay procedures to increase multiplexing levels from 48-plex to 1,152-plex within a period of about 12 months (Fig. 3). Since

Principal Investigator **Chee, Mark S.**

Proprietary Info

Proprietary Info

Proprietary Info

Examples of data at the individual bead level are shown in Fig.4. The data quality is high. The data in each panel are from 96 DNAs (i.e. each analyzed on a different array, each at 1,152-plex throughout the genotyping process).

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

C.5 Assay Development Capacity

Proprietary Info

Proprietary Info

Our assay development process is managed by the genotyping LIMS and is highly scalable. Custom assay design software is used to generate oligonucleotide sequences. These sequences are sent to our Oligator™ facility (described in the following section) for synthesis and also stored in the genotyping LIMS. The Oligator facility delivers oligonucleotides to the genotyping production facility in 96-well plates. Incoming plates containing individual oligonucleotides are identified and tracked by LIMS, which also generates pooling instructions. Robots pool the oligonucleotides under LIMS supervision. The pools are stored in barcoded microtiter plates and accessioned for production genotyping using the LIMS. The entire process is designed to ensure accuracy and scalability.

Proprietary Info

Proprietary Info

Proprietary Info

C.6 Oligonucleotide Synthesis

To enable cost-effective assay development, oligonucleotides are manufactured at Illumina using a proprietary high-throughput DNA synthesizer called the Oligator™ (Fig. 5). Proprietary Info

Proprietary Info

Proprietary Info

This results in substantial savings in labor.

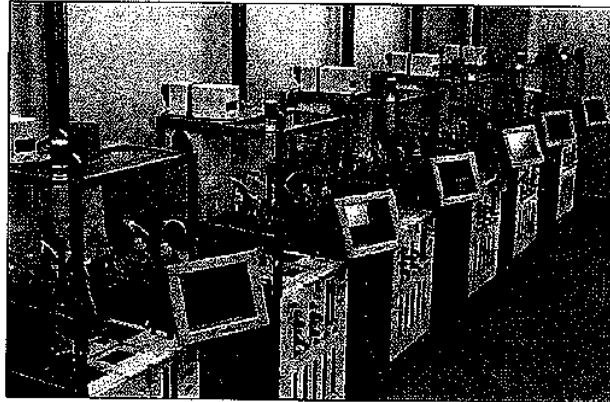
Proprietary Info

Proprietary Info

Proprietary Info

Oligonucleotides are quality-controlled by mass spectrometry.

Principal Investigator Chee, Mark S.



Proprietary Info

C.7 Costs

Proprietary Info

C.8 Accountability and Prior Experience

As a commercial organization, we are accountable to our customers through genotyping contracts that specify performance metrics and turnaround times. In addition, we have prior experience in completing large-scale genomics projects.

Proprietary Info

Proprietary Info

each of the processes and equipment.

Proprietary Info

A LIMS was used to track all the samples through

Proprietary Info

D. RESEARCH DESIGN AND METHODS

D.1 Genotyping Technology and System

We have designed and developed a large-scale genotyping system essentially from the ground up

Proprietary Info

Propriet
Info

Although key aspects of the system are proprietary, access to the technology is readily available. Illumina provides fee-for-service genotyping that may include assay development, and is also developing systems for commercial release in the second half of 2002.

An overall description of the genotyping production system together with analyses of performance is provided in Section C

Proprietary Info

Proprietary Info

D.1.1 Assay Format is Designed for Multiplexing

Proprietary Info

Proprietary Info

The first step in conventional SNP genotyping is to amplify the SNP of interest from genomic DNA. In contrast, we perform an allelic discrimination step directly on genomic DNA. This is done using two allele-specific oligos, each 5' tailed with a different universal PCR priming sequence (Fig. 6). The product of this enzymatic procedure is an allele-specific PCR template. (In the case of a heterozygous DNA, both allele-specific PCR templates are produced at this step). PCR is then carried out, using only three primers (P1', P2' and P3' in Fig. 6). We have found that this procedure is amenable to high levels of multiplexing. Currently, when sufficient SNPs are available, we multiplex up to 1,152 genotyping reactions

Proprietary Info

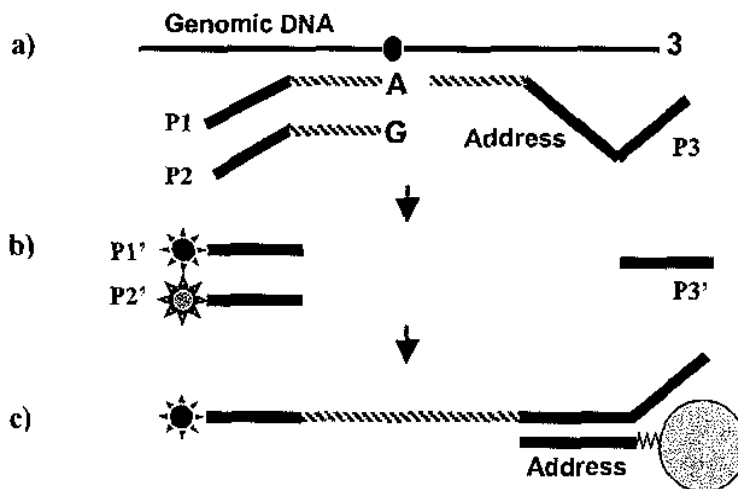


Figure 6. Assay format. a) For each SNP of interest, two allele-specific oligos and a locus-specific common oligo are annealed to genomic DNA (e.g. in a 1,152-plex reaction, a total of 3,456 oligos are annealed simultaneously, in the same reaction well in a microtiter plate). If an allele-specific oligonucleotide is complementary to the genomic DNA, a ligation product is formed. This product has universal priming sites at the 5' and 3' ends (i.e. P1 or P2, and P3). If the genomic DNA is heterozygous, then two products are formed: P1-P3 and P2-P3. b) Universal primers are added and PCR is carried out. The two allele-specific universal

primers, P1' and P2' are fluorescently labeled, each with a different dye. Each amplicon contains an address that is complementary to a probe in the array, so that the genotype of each SNP can be read out on a different bead type in the array. c) The PCR amplicons are hybridized to an array of beads. The ratio of the two fluorescent signals indicates the genotype.

Another key aspect of the assay design is the incorporation of an address sequence, so that the assay products can be read out on a universal array (Chen et al. 2000; Fan et al. 2000; Gerry et al. 1999; Iannone et al. 2000). This provides flexibility. The probes on the array are random, artificial sequences that are not SNP-specific. Any set of SNPs can be analyzed simply by building the address sequences into the SNP-specific assay oligonucleotides (Section D.1.1).

The use of a universal array simplifies manufacturing and reduces costs. The universal array is implemented on our BeadArray™ platform, detailed below.

D.1.2 Array Matrix Platform

The randomly ordered BeadArray technology, invented at Tufts University (Michael et al. 1998; Walt 2000), has been developed at Illumina as a platform for SNP genotyping and other high-throughput assays. Each array is assembled on an optical imaging fiber bundle consisting of about 50,000 individual fibers fused together into a hexagonally packed matrix. The ends of the bundle are polished, and one end is etched to produce a well in each fiber. This process takes advantage of the intrinsic structure of the optical fibers in the bundle (Fig. 7).

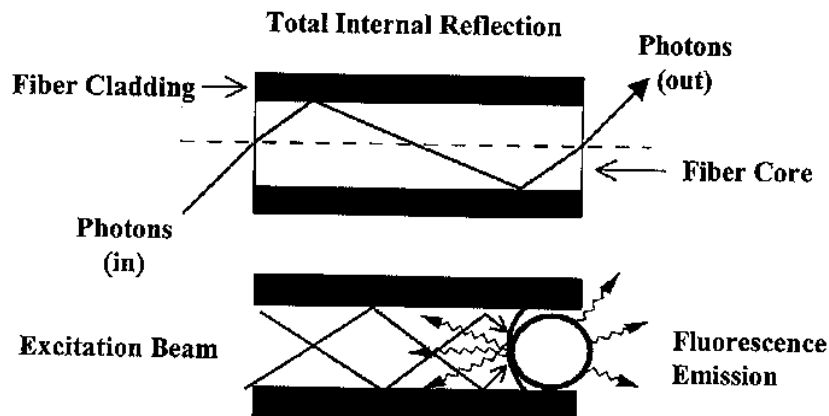


Figure 7. Structure and useful properties of an individual optical fiber. Each fiber has a light-conducting inner core that is surrounded by a cladding of different refractive index. The core can be chemically etched at a different rate from its surrounding cladding. By treating the polished end of an optical fiber with acid, an array of microwells is generated. The geometry and dimensions of the array are determined by the physical specifications of the optical fiber, and are chosen so that one bead can fit in each well in the array. Once a labeled target nucleic acid is hybridized to beads in the array, a fluorescent signal can be generated by making use of the optical properties of the fiber. An excitation beam is guided to the bead through the fiber bundle, and emitted fluorescence is guided back up the fiber, allow the array to be imaged at the opposite end of the optical fiber bundle.

After polishing and etching, a fiber bundle can hold up to 50,000 beads, each ~3 microns in diameter and spaced ~5 microns apart. This highly miniaturized array is about 1.4 mm across and has a packing density of 40,000 array elements per square millimeter – approximately 400 times the information density of a typical spotted microarray with 100 μm spacing. Each derivatized bead has several hundred thousand copies of a particular oligonucleotide covalently attached. Bead libraries are prepared by conjugation of

oligonucleotides to silica beads, followed by quantitative pooling together of the individual bead types. The preparation of a bead library and assembly into an array are illustrated in Fig. 8. The raw materials are minimal and most of the key manufacturing steps are bulk processing steps. These factors result in low manufacturing costs and highly reproducible arrays.

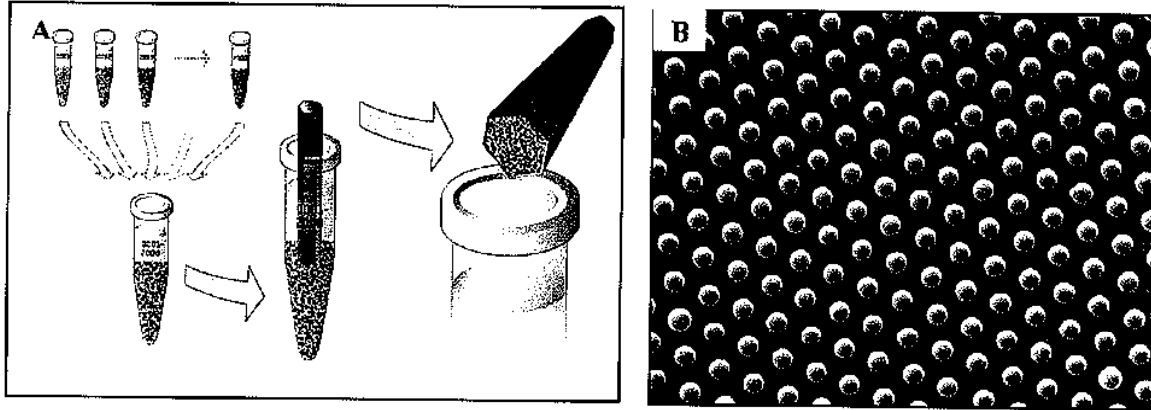


Figure 8. Assembly of a randomly ordered fiber optic array. (A) A collection of bead types, each with a distinct oligonucleotide capture probe, is pooled. An etched fiber optic bundle is dipped into the bead pool, allowing individual beads to assemble into the microwells at the bundle's end. **(B)** Scanning electron micrograph of an assembled array containing 3 micron diameter silica beads. The beads are stably associated with the wells under standard hybridization conditions.

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

Principal Investigator **Chee, Mark S.**

Proprietary Info



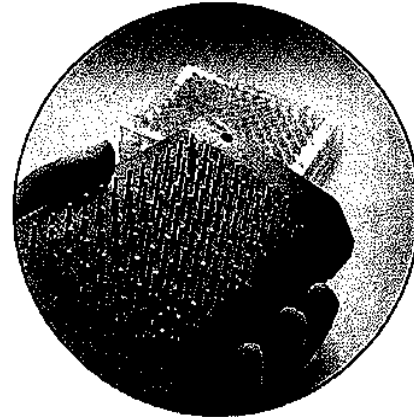
Proprietary Info

The increase in numerical precision provided by redundancy is also helpful in generating a meaningful quality (GenCall) score.

D.1.5 The Array Matrix Format is Designed for Parallel Sample Processing

To further increase throughput, the arrays are formatted into a matrix, in a pattern that matches the wells of standard microtiter plates (Fig. 10).

Figure 10. Photograph of a 96-array matrix. Each array is located on the end of an optical fiber bundle containing ~ 50,000 individual fibers. The spacing of the arrays matches that of a 96-well plate, allowing 96 separate samples to be processed simultaneously.



The matrix format allows streamlined sample handling. By matching the spacing to that of a standard 96-well microtiter plate, off-the-shelf technologies can be used for automated pipetting, plate handling, and positive tracking with barcodes. Thus, the microtiter format provides a straightforward and highly adaptable path to parallel sample processing. In contrast, most conventional array systems are not well suited to the efficient processing of many samples.

D.2 Genotype Production Plan

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

Proprietary Info

d)

Proprietary Info

D.2.1 Process Refinement

Proprietary Info

Proprietary Info

Proprietary Info

D.2.2 Strategies for Cost Reduction

Proprietary Info

Principal Investigator **Chee, Mark S.**

Proprietary Info

Proprietary Info

D.2.3 Production Genotyping

All genotyping will be carried out in Illumina's large-scale, positively-tracked, genotyping production facility described in Section C. Quality will be monitored continuously using LIMS to track metrics and GenCall scores to production processes and objects. Each sample contains a set of internal controls that is read out on the array, including controls for hybridization, allele-specificity, and balanced amplification of both alleles. Each set of 96 samples processed contains control DNAs.

Proprietary Info

The amount of genotyping proposed is well within the capacity of the genotyping facility, which is easily scaled to higher capacities.

Proprietary Info

Proprietary Info

D.2.3.1 Production Schedule

In the first 6 months of the project we will refine processes, explore cost reduction options, and develop assays for and genotype 40,000 SNPs. We will also put in place our data submission procedures and other protocols needed for the project.

Proprietary Info

Proprietary Info

D.3 Project Cost Analysis

Proprietary Info

D.3.1

Proprietary Info

Proprietary Info

D.3.2

Proprietary Info

Proprietary Info

Principal Investigator Chee, Mark S.

Proprietary Info

D.3.3

Proprietary Info

Proprietary Info

D.3.4

Proprietary Info

Proprietary Info

It is also possible to vary the number of SNPs analyzed.

Proprietary Info

Proprietary Info

Illumina operates as a business in a competitive field, and is therefore not prepared to provide a full breakdown of genotyping and oligonucleotide costs. However, the budget for this project is not subsidized in any way – i.e. the generation of genotypes is fully paid for by the requested budget. Our costs are monitored through an enterprise resource planning (ERP) system which tracks and inventories materials and accounts for labor.

Proprietary Info

D.4 Obtaining Additional SNPs in Regions

While we expect that most of the SNPs needed for this project will be available, we will also work with our collaborators to develop a strategy for obtaining SNPs in gaps. Towards the end of the project, it will likely become efficient to carry out some directed resequencing from PCR amplicons. Illumina currently makes PCR primers and pools them in pairs, in 96-well microtiter plates. We may be able to coordinate with our collaborators at the Sanger and Whitehead Institutes to automate the design and synthesis of primer pairs for finding SNPs for gap filling, and supply them with low-cost primer pairs for this purpose.

D.5 Data Analysis

D.5.1 Finding Blocks and Haplotypes

Proprietary Info

Proprietary Info

These tools and the expertise gained during the collaboration will be used in

the present project.

Proprietary Info

Proprietary Info

One element of particular strength in the present project is Illumina's development of a GenCall score for each genotype.

Proprietary Info

Proprietary Info

Proprietary Info

(Abecasis et al. 2002).

Proprietary Info

Proprietary Info

Proprietary Info

D.5.2 Choosing Tag SNPs

Proprietary Info

D.5.3 Making Genotype and Haplotype Data Publicly Available

Genotyping at Illumina is carried out in a massively parallel way. Each round of SNP assay development will be treated by the genotyping production operation as a separate project, with all the data from that round becoming available at once. As each round is completed and the data have been checked

for quality, we will submit the genotypes electronically to one or more data repositories agreed upon by the Coordinating Committee. In advance of the first submission, we will work with the recipients to ensure that our data format can be automatically and accurately downloaded. Each genotype will be provided together with its GenCall score. Similarly, we will work with recipient(s) in advance to determine the best way to submit haplotype data as it becomes available. We will follow the release policies and procedures agreed to by the Coordinating Committee.

D.6 Management Plan

The management of the project will be carried out by the PI and key personnel at Illumina. This team will define and generate progress reports for internal use, and the PI will submit periodic progress reports in a standard format as agreed upon by the Coordinating Committee and the Scientific Advisory Panel. The team will meet frequently and regularly and monitor key metrics to ensure that the project is on track.

Proprietary Info

Proprietary Info

Proprietary Info

Dr. Chee will carry out the PI's responsibilities as defined in the HapMap

RFA.

Proprietary Info

Proprietary Info

Proprietary Info

E. HUMAN SUBJECTS

Proprietary Info

Proprietary Info

Proprietary Info

Thus, while the research funded under this RFA will involve human subjects as defined in Title 45 CFR, Part 46, an exemption should be recognized under **Section 46.101(b)(4).**

Proprietary Info

Proprietary Info

Inclusion of Women: Roughly equal numbers of females and males will be studied.

Proprietary Info

Proprietary Info

F. VERTEBRATE ANIMALS

Proprietary Info

G. LITERATURE CITED

- Abecasis, G.R., S.S. Cherny, W.O. Cookson, and L.R. Cardon. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**: 97-101.
- Chen, J., M.A. Iannone, M.S. Li, J.D. Taylor, P. Rivers, A.J. Nelsen, K.A. Slentz-Kesler, A. Roses, and M.P. Weiner. 2000. A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res* **10**: 549-557.
- Daly, M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* **29**: 229-232.
- Dawson, E. et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*: In press.
- Deloukas, P. L.H. Matthews J. Ashurst J. Burton J.G. Gilbert M. Jones G. Stavrides J.P. Almeida A.K. Babbage C.L. Bagguley J. Bailey K.F. Barlow K.N. Bates L.M. Beard D.M. Beare O.P. Beasley C.P. Bird S.E. Blakey A.M. Bridgeman A.J. Brown D. Buck W. Burrill A.P. Butler C. Carder N.P. Carter J.C. Chapman M. Clamp G. Clark L.N. Clark S.Y. Clark C.M. Clee S. Clegg V.E. Copley R.E. Collier R. Connor N.R. Corby A. Coulson G.J. Coville R. Deadman P. Dhani M. Dunn A.G. Ellington J.A. Frankland A. Fraser L. French P. Garner D.V. Grafham C. Griffiths M.N. Griffiths R. Gwilliam R.E. Hall S. Hammond J.L. Harley P.D. Heath S. Ho J.L. Holden P.J. Howden E. Huckie A.R. Hunt S.E. Hunt K. Jekosch C.M. Johnson D. Johnson M.P. Kay A.M. Kimberley A. King A. Knights G.K. Laird S. Lawlor M.H. Lehtvaslaiho M. Leversha C. Lloyd D.M. Lloyd J.D. Lovell V.L. Marsh S.L. Martin L.J. McConnell K. McLay A.A. McMurray S. Milne D. Mistry M.J. Moore J.C. Mullikin T. Nickerson K. Oliver A. Parker R. Patel T.A. Pearce A.I. Peck B.J. Phillimore S.R. Prathalingam R.W. Plumb H. Ramsay C.M. Rice M.T. Ross C.E. Scott H.K. Sehra R. Shownkeen S.

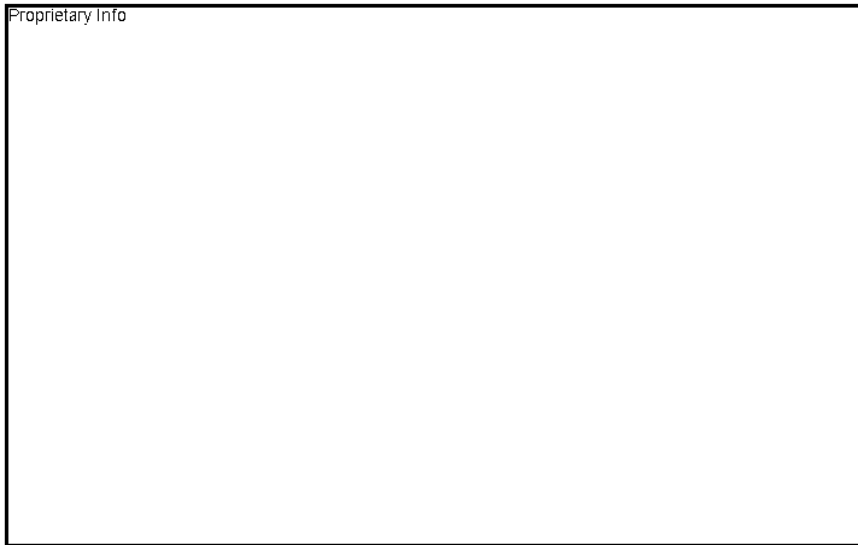
- Sims C.D. Skuce M.L. Smith C. Soderlund C.A. Steward J.E. Sulston M. Swann N. Sycamore R. Taylor L. Tee D.W. Thomas A. Thorpe A. Tracey A.C. Tromans M. Vaudin M. Wall J.M. Wallis S.L. Whitehead P. Whittaker D.L. Willey L. Williams S.A. Williams L. Wilming P.W. Wray T. Hubbard R.M. Durbin D.R. Bentley S. Beck and J. Rogers. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**: 865-871.
- Fan, J.B., X. Chen, M.K. Halushka, A. Berno, X. Huang, T. Ryder, R.J. Lipshutz, D.J. Lockhart, and A. Chakravarti. 2000. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res* **10**: 853-860.
- Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The Structure of Haplotype Blocks in the Human Genome. *Science*.
- Gerry, N.P., N.E. Witowski, J. Day, R.P. Hammer, G. Barany, and F. Barany. 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J Mol Biol* **292**: 251-262.
- Iannone, M.A., J.D. Taylor, J. Chen, M.S. Li, P. Rivers, K.A. Slentz-Kesler, and M.P. Weiner. 2000. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* **39**: 131-140.
- Jeffreys, A.J., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**: 217-222.
- Johnson, G.C., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C. Gough, D.G. Clayton, and J.A. Todd. 2001. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**: 233-237.
- Kwok, P.Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**: 235-258.
- Michael, K.L., L.C. Taylor, S.L. Schultz, and D.R. Walt. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal Chem* **70**: 1242-1248.
- Patil, N., A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K.A. Frazer, S.P. Fodor, and D.R. Cox. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719-1723.
- Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411**: 199-204.
- Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers,

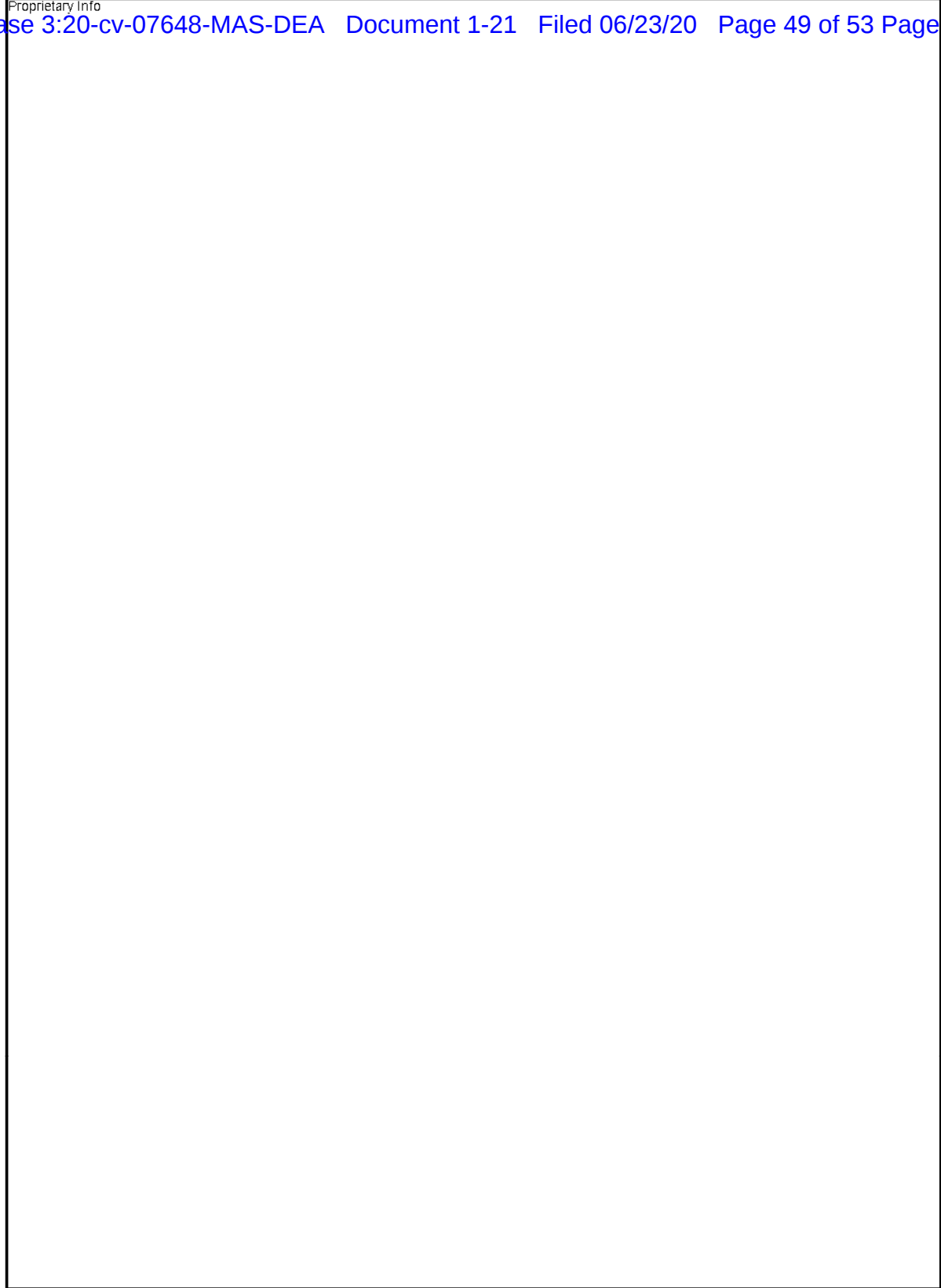
D.R. Bentley, P.Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, and D. Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

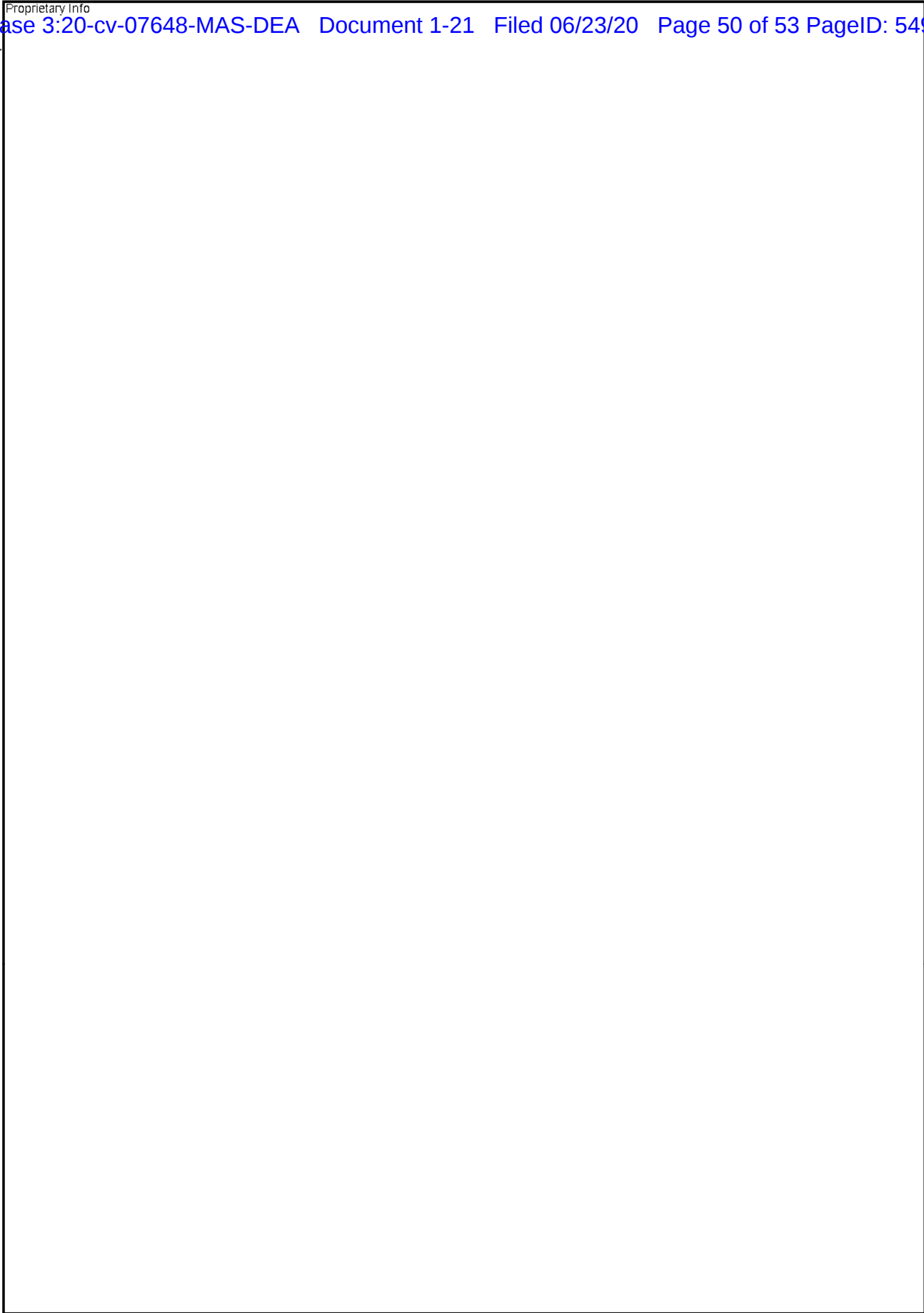
Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978-989.

Syvanen, A.C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**: 930-942.

Walt, D.R. 2000. Techview: molecular biology. Bead-based fiber-optic arrays. *Science* **287**: 451-452.








Illumina, Inc.
Overhead Rate Calculator

Proprietary Info



Illumina, Inc.
Overhead Rate Calculator

Proprietary Info



Principal Investigator/Program Director (Last, first, middle):

CHECKLIST

TYPE OF APPLICATION (Check all that apply.)

- NEW application.** (This application is being submitted to the PHS for the first time.)
 - SBIR Phase I SBIR Phase II: SBIR Phase I Grant No. _____ SBIR Fast Track
 - STTR Phase I STTR Phase II: STTR Phase I Grant No. _____ STTR Fast Track

- REVISION** of application number: _____
(This application replaces a prior unfunded version of a new, competing continuation, or supplemental application.)
- COMPETING CONTINUATION** of grant number: _____
(This application is to extend a funded grant beyond its current project period.)
- SUPPLEMENT** to grant number: _____
(This application is for additional funds to supplement a currently funded grant.)

- INVENTIONS AND PATENTS**
(Competing continuation appl. and Phase II only)
- No
 - Previously reported
 - Yes. If "Yes," Not previously reported

- CHANGE** of principal investigator/program director.
Name of former principal investigator/program director: _____

- FOREIGN** application or significant foreign component.

1. PROGRAM INCOME (See instructions.)

All applications must indicate whether program income is anticipated during the period(s) for which grant support is request. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

2. ASSURANCES/CERTIFICATIONS (See instructions.)

The following assurances/certifications are made and verified by the signature of the Official Signing for Applicant Organization on the Face Page of the application. Descriptions of individual assurances/certifications are provided in Section III. If unable to certify compliance, where applicable, provide an explanation and place it after this page.

- Human Subjects; •Research Using Human Embryonic Stem Cells•
- Research on Transplantation of Human Fetal Tissue •Women and Minority Inclusion Policy •Inclusion of Children Policy• Vertebrate Animals•

- Debarment and Suspension; •Drug- Free Workplace (applicable to new [Type 1] or revised [Type 1] applications only); •Lobbying; •Non-Delinquency on Federal Debt; •Research Misconduct; •Civil Rights (Form HHS 441 or HHS 690); •Handicapped Individuals (Form HHS 641 or HHS 690); •Sex Discrimination (Form HHS 639-A or HHS 690); •Age Discrimination (Form HHS 680 or HHS 690); •Recombinant DNA and Human Gene Transfer Research; •Financial Conflict of Interest (except Phase I SBIR/STTR) •STTR ONLY: Certification of Research Institution Participation.

3. FACILITIES AND ADMINISTRATIVE COSTS (F&A)/ INDIRECT COSTS. See specific instructions.

- DHHS Agreement dated: 6/8/2000 No Facilities And Administrative Costs Requested.
- DHHS Agreement being negotiated with _____ Regional Office.
- No DHHS Agreement, but rate established with _____ Date _____

CALCULATION* (The entire grant application, including the Checklist, will be reproduced and provided to peer reviewers as confidential information.)

a. Initial budget period:	Amount of base \$ _____	x Rate applied _____	% = F&A costs _____	\$ _____
b. 02 year	Amount of base \$ _____	x Rate applied _____	% = F&A costs _____	\$ _____
c. 03 year	Amount of base \$ _____	x Rate applied _____	% = F&A costs _____	\$ _____
d. 04 year	Amount of base \$ _____	x Rate applied _____	% = F&A costs _____	\$ _____
e. 05 year	Amount of base \$ _____	x Rate applied _____	% = F&A costs _____	\$ _____
TOTAL F&A Costs \$				<input style="width: 100px; height: 20px;" type="text"/>

*Check appropriate box(es):

- Salary and wages base Modified total direct cost base Other base (Explain)
- Off-site, other special rate, or more than one rate involved (Explain)

Explanation (Attach separate sheet, if necessary.): **Attached**

4. SMOKE-FREE WORKPLACE Yes No (The response to this question has no impact on the review or funding of this application.)