

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

QIAGEN Sciences, LLC,
Petitioner,

v.

Tecan Genomics, Inc.,
Patent Owner

Case IPR2025-00028
Case IPR2025-00029
U.S. Patent Nos. 10,036,012 and 10,876,108

DECLARATION OF CARLOS D. BUSTAMANTE, PH.D.

TABLE OF CONTENTS

I.	BACKGROUND AND QUALIFICATIONS	4
II.	MATERIALS CONSIDERED	9
III.	LEGAL STANDARDS	10
IV.	TECHNICAL BACKGROUND	13
A.	Genetic Information	13
B.	Nucleic Acid Genotyping	16
C.	Library Preparation.....	18
D.	Anchored PCR and Genome Walking.....	23
E.	PCR-Free Sequencing Methods	33
F.	Hybridization Arrays	35
G.	Nucleic Acid Sequencing	43
1.	Pre-NGS Technologies for Sequencing.....	43
2.	Next Generation Sequencing.....	46
H.	Target Enrichment	51
1.	PCR-based Target Enrichment Methods	52
2.	Capture-by-Hybridization	58
3.	Capture-by-Circularization.....	59
I.	DNA Hybridization Microarrays and NGS Platforms Rely on Different Technology.....	61
V.	THE PERSON OF ORDINARY SKILL IN THE ART	70
VI.	THE PATENTED INVENTIONS	71
A.	The Claims of the '012 Patent.....	72
B.	The Claims of the '108 Patent.....	76
C.	Illustration of the Claimed Inventions.....	79
D.	Relevant Prosecution History.....	88
E.	CLAIM CONSTRUCTION	91
VII.	RELEVANT PRIOR ART	92
A.	Jones (EX1011)	92
B.	Shapero (EX1006).....	95

1. The MARA Method.....	98
2. Shapero Validates MARA Using Genotyping Array Results.....	103
C. Syvänen (EX1012)	110
D. Delseny (EX1007).....	113
1. Delseny Does Not Discuss the Use of a Single Target-Specific Primer for Targeted Sequencing.....	114
2. Delseny’s Discussion of Using NGS for SNP Genotyping Concerns Whole Genome Sequencing	120
E. Han 2009 (EX2117)	122
VIII. Neither Shapero Nor Delseny Disclose All Elements of the Independent Claims	127
IX. A POSA Would Not Have Been Motivated to Combine Shapero with Delseny With a Reasonable Expectation of Success	131
A. A POSA Would Not Have Been Motivated to Adapt a Method Used with Hybridization Arrays for Use With NGS In the Manner Claimed by the Patents	131
B. Shapero Fails to Resolve the Concerns with Target Specificity and Off-Target Binding.....	136
C. A POSA Reviewing Shapero Would Not Have Discerned a Need to Switch from Hybridization Arrays to NGS.....	145
D. A POSA Would Not Have Understood Shapero’s Reference to Syvänen to Teach that the MARA Method Could be Used with Sequencing Technologies.	148
E. A POSA Would Not Have Been Motivated by Delseny to Adapt the MARA Method for NGS	151
F. A POSA Would Not Have Been Motivated to Modify the MARA Method to Practice NGS Given the Weight of the Prior Art.....	157
G. Dr. Metzker’s Prior Publication Would Have Discouraged a POSA from Modifying Shapero to Practice the Claimed Inventions.....	161
H. There is No Reason a POSA Would Have Decided to Modify Shapero to Practice the Claimed Methods Instead of Other Target Capture Approaches	164
I. A POSA Would Not Have Had a Reasonable Expectation of Success in Modifying the MARA Method to Practice NGS As Practiced by the Claimed Inventions.....	165
1. Shapero’s Call Rates and DS Are Inapplicable.....	166

2. Other Data Reported by Shapero Would Have Diminished a POSA's Already Low Expectations	168
3. Delseny's Disclosure on Using Genotyping with NGS is Inapplicable....	169
J. How the Art Cited to Shapero Demonstrates the Limited Utility of MARA 170	
X. SECONDARY CONSIDERATIONS.....	186
A. Nexus.....	187
B. Teach-Away/Skepticism	192
C. Long-felt, unmet need	193
D. Industry Praise for Tecan's Solution to these Long-Felt, Unmet Needs...	197
E. QIAGEN'S SPE Products Are Copies of Tecan's SPET Products	200
F. QIAGEN's Prior Statements Concerning its SPE Products.....	205
G. QIAGEN Previously Demonstrated Unexpected Results Using the Claimed Inventions.....	212
XI. CONCLUSION	219

I, Carlos D. Bustamante, declare as follows:

1. I have been retained by Patent Owner, Tecan Genomics, Inc. (“Tecan”) as an expert witness in this *inter partes* review.

2. As part of my assignment, I have been asked to review Petitioner QIAGEN Sciences, LLC (“QIAGEN”)’s petitions for *inter partes* review and to provide my opinions regarding positions taken by QIAGEN.¹

3. I previously submitted a declaration as part of Patent Owner’s Preliminary Response. *See* EX2028. In that declaration, I considered QIAGEN’s allegations that a person of ordinary skill in the art would consider claim 1 of U.S. Patent No. 10,036,012 (“the ’012 patent”) and claim 1 of U.S. Patent No 10,876,108

¹ In reviewing the petitions and the accompanying declarations from Dr. Metzker, I have concluded that the documents do not substantively differ with respect to their discussions of the independent claims of the challenged patents. For simplicity, when I refer to the “Petition,” I am referring to the petition filed in IPR2025-00028. Similarly, when I cite to the declaration of Dr. Metzker (EX1003), I am referring to his petition submitted in IPR2025-00028. In both instances, my references refer equally to the analogous (and near identical) documents submitted in IPR2025-00029, which do not materially differ with respect to the disclosures identified herein.

(“the ’108 patent”) to be obvious over Shapero (EX1006) in view of Delseny (EX1007). Because I determined that a POSA would not have viewed the independent claims to be obvious over Shapero and Delseny, I concluded that a POSA would find the challenged dependent claims of the ’012 and ’108 patents to be nonobvious for at least the same reasons. My opinions remain the same.

4. I understand that a panel from the United States Patent and Trademark Office’s Patent Trial and Appeal Board (“PTAB”) instituted *inter partes* review of the ’012 and ’108 patents. I have reviewed those institution decisions. This declaration expands on my prior testimony, addresses the issues raised in the institution decisions, and considers additional issues with QIAGEN’s claims. Where helpful, I have repeated relevant concepts discussed in EX2028.

5. In assessing obviousness, I have considered the contents of the ’012 and ’108 patents, the file histories of the patents and related patents, prior art publications, the two petitions for *inter partes* review, and the Declaration of Michael L. Metzker, Ph.D. (EX1003)—a declaration that QIAGEN relies on to support its claims. I have considered the prior art relied on by QIAGEN in its Petition: Shapero (EX1006), Delseny (EX1007), Hamady (EX1033), and Jones (EX1011). I have also relied on my own experiences as a person of ordinary skill in the art as of the effective filing date of the ’012 and ’108 patents (January 26, 2012).

6. I have considered QIAGEN's allegations that the challenged claims are rendered obvious by Shapero and Delseny (Ground 1). As explained further below, I disagree with QIAGEN's assessment. Shapero and Delseny do not render obvious the challenged claims of the '012 and '108 patents. In large part, QIAGEN's petitions are biased by hindsight, arguing that a POSA would have been motivated to adapt a preparation method for hybridization arrays to instead be used with PCR-based target capture and Next Generation Sequencing ("NGS") even though both suffered particular sensitivities that were not of the same concern for hybridization arrays. In fact, in 2012 there was broad agreement that hybridization arrays were generally highly accurate and reproducible for large scale genetic analysis of SNPs. In contrast, there was a lack of standardization for other methods, especially NGS sequencing where target specificity and off-target binding posed particular challenges. QIAGEN, however, dismisses the concerns in the art as of 2012 and asserts that Shapero's results—which are based on the use of hybridization arrays and not NGS—somehow disprove the conventional wisdom of the time. To the contrary, these concerns would have discouraged a POSA from modifying the MARA method disclosed in Jones and Shapero to practice the claimed methods. A POSA would have lacked a reasonable expectation of success given the concerns in the art and the constraints of Shapero. If anything, Shapero and Delseny would have only exacerbated a POSA's discouragement and expectations of failure.

7. I have also been asked to consider whether any objective indicia of nonobviousness apply to the claimed inventions. I have concluded that at least several objective indicia are applicable and that all support the nonobviousness of the claimed inventions, which are embodied in Tecan's SPET products, including Tecan's Allegro® Targeted Genotyping and Ovation Target Enrichment products.

I. BACKGROUND AND QUALIFICATIONS

8. In this section, I summarize my educational background, career history, publications, and other relevant qualifications.

9. I have extensive experience in the interface of computational and biological sciences, as well as its application in medicine, agriculture, and evolutionary biology, spanning a career of over 25 years in the field. My education, work experiences, and research have focused on numerous relevant topics, including genomic sequencing and analysis, sample and library preparation for genomic sequencing, biostatistics, and genotype assay design. My expertise in these areas inform my opinions herein.

10. In 1997, I graduated magna cum laude, highest honors, with a Bachelor of Arts in Biology from Harvard University. Then, in 2001, I earned both my M.S. in Statistics and Ph.D. in Biology from Harvard University. From 2001 to 2002, I served as a postdoctoral researcher in mathematical genetics at the University of Oxford funded by a Marshall-Sherfield Scholarship.

11. I am currently an adjunct professor of Biomedical Data Science, Genetics, and (by courtesy) Biology at Stanford University where I have served since 2010. I also served as the Inaugural Chair of the Department of Biomedical Data Sciences from 2015-2019.

12. I am a named inventor on numerous relevant patents and patent publications, including: U.S. Publication No. 2015/0360194 (“*Enrichment of DNA Sequencing Libraries from Samples Containing Small Amounts of Target DNA*”); U.S. Publication No. 2015/0232834 (“*Partitioning of DNA Sequencing Libraries into Host and Microbial Components*”); and U.S. Publication No. 2014/0045705 (“*Techniques for Determining Haplotype by Population Genotype and Sequence Data*”)—all of which are now issued patents. I have also completed numerous projects with the NIH, such as: *Inferring Multiple-SNP Disease Association with DNA Resequencing Data* (NIH/NIMH, 2008-2012); *Population Genetic Inferences from Dense Genotype Data* (NIH/NHGRI, 2010-2013); and *Methods for High-Resolution Analysis of Genetic Effects on Gene Expression* (NIH, 2013-2016).

13. In 2002, I joined Cornell University as an Assistant Professor in the Department of Biological Statistics and Computational Biology and Department of Statistical Sciences. I later became an Associate and Full tenured Professor at Cornell University, and also served as the Associate Director of the Cornell Center

for Comparative and Population Genomics. At Cornell, my work primarily focused on population genetics and agricultural genomics.

14. I am the founder of Galatea Bio, Inc. and serve as President, CEO, and Board Member of the company. I served as a venture partner at Fidelity's early stage biotechnology group, F-Prime Capital, where the company was incubated during my tenure, and retain a venture partner title while serving as an operator of Galatea Bio. I am also the Principal and Chairman of CDB Consulting Group, Inc., an internationally recognized consulting firm, providing data science and genomics expertise in the areas of medicine, agriculture, and biology. I serve as a Board observer at Embark Veterinary, a company I advised at CDB Consulting and where F-Prime Capital is an investor.

15. I am also the founder of Arc-Bio (formerly IdentifyGenomics and BigData Bio) and now part of Cantata Bio. I served as chair of its Scientific Advisory Board and board member until the company was merged into Cantata Bio. Further, I was on the Board of Directors of Etalon Diagnostics and Eden Roc Sciences, two companies I helped start.

16. I have served on the advisory boards of various other corporations, non-profit entities, and government organizations. For example, I am currently a member of the scientific advisory boards at the following companies: Fauna Bio and Cantata Bio. I was also a member of the National Advisory Council to the National Human

Genome Research Institute at the National Institutes of Health and remain on the Scientific Advisory Board of NHGRI Center for Research on Genomics and Global Health. I was also the Co-Chair of the PAGE Analysis Working Group, the ClinGen Ancestry and Diversity Working Group, and the ClinGen Complex Disease/Polygenic Risk Score working group.

17. I have served on the Editorial Boards of numerous publications in the field, including PLoS-Genetics, American Journal of Human Genetics, Human Biology, Genome Research, Molecular Biology and Evolution, and Bioinformatics.

18. I have authored or co-authored over 250 articles and publications, including: *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes*, 337 SCIENCE 6090 (2012); *Native American gene flow into Polynesia predating Easter Island settlement.*, 583(7817) NATURE 572-577 (2020); *Deconvoluting complex correlates of COVID-19 severity with a multi-omic pandemic tracking strategy*, NAT COMMUN (2022); *Polygenic risk score portability for common diseases across genetically diverse populations*, 18(1) HUM GENOMICS 93 (2024); *Neural ADMIXTURE for rapid genomic clustering*, 7 NAT COMPUT SCI. 621-629 (2023); *Validating and automating learning of cardiometabolic polygenic risk scores from direct-to-consumer genetic and phenotypic data: implications for scaling precision health research*, 16(1) HUM GENOMICS 37 (2022); *Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations*,

107(4) AM J HUM GENET. 788-789 (2020); *Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing*, 4 G3 1681 (2014); *Achieving High-Sensitivity for Clinical Applications Using Augmented Exome Sequencing*, 7 GENOME MED. 71 (2015); *A Research Roadmap for Next-Generation Sequencing Informatics*, 8 SCI. TRANSL. MED. 335 (2016); *Data Mining of Digitized Health Records in a Resource-Constrained Setting Reveals that Timely Immunophenotyping is Associated with Improved Breast Cancer Outcomes*, 18(1) BMC CANCER 933 (2018); *Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits*, 570 NATURE 514 (2019); *Structural Variation Detection by Proximity Ligation from Formalin-Fixed, Paraffin-Embedded Tumor Tissue*, 21(3) J. MOLECULAR DIAGNOSTICS 375 (2019), *Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males versus Females*, 341 SCIENCE 562 (2013); and *Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences*, 48 NAT GENET. 593 (2016).

19. For my work in the field, I have received several honors and fellowships. These honors and fellowships include: the MacArthur “Genius” Award (2010), a Marshall-Sherfield Scholarship to Oxford (2001-2002), the Stanford Prize in Population Genetics and Society (2016); the Cornell Provost Award for Distinguished Research (2008); a Sloan Research Fellowship in Molecular Biology

(2007-2009); and a Howard Hughes Medical Institute Graduate Research Fellowship (1999-2001).

20. A copy of my curriculum vitae, attached as **Exhibit A**, contains further details on my education, experience, publications, and the other qualifications allowing me to render an expert opinion in this matter.

21. I am being compensated my standard rate of \$800 per hour, plus expenses. My compensation is in no way dependent on the content of my testimony or upon the outcome of this proceeding.

22. In the past five years, I have testified as an expert witness in five other matters: (1) *Foundation Medicine Inc v. Guardant Health, Inc.*, IPR2019-00634 (P.T.A.B.); (2) *TwinStrand Biosciences, Inc. v. Guardant Health, Inc.*, C. A. 21-1126-GBW-SRF (D. Del.); (3) *State v Pease*, 88-10 (D. Mont.); (4) *Integrated DNA Technologies, Inc. v. Tecan Genomics, Inc.*, IPR2025-00015, IPR2025-00016 (P.T.A.B.); and (5) *Integrated DNA Technologies, Inc. v. Tecan Genomics, Inc.*, IPR2024-01504 (P.T.A.B.).

II. MATERIALS CONSIDERED

23. In formulating the opinions in this declaration, I have relied on over 25 years of work experience and knowledge regarding bioinformatics, genomics,

nucleic acid sequencing, and hybridization array design and associated analysis/interpretation. I also considered the documents cited in this declaration.

III. LEGAL STANDARDS

24. I have been informed by counsel for Tecan of certain legal standards, and I have applied these standards in rendering my opinions.

25. When I say in this Declaration “I understand,” I mean that “I understand from counsel for Tecan in this IPR.”

26. I understand that to prove obviousness in an *inter partes* review, it is the petitioner’s ultimate burden to prove obviousness by a preponderance of the evidence.

27. I understand that the question of whether the claims of a patent are obvious is considered from the perspective of a person of ordinary skill in the art (the “POSA”).

28. I understand that the POSA is a legal construct representing a hypothetical individual with the same level of skill as the typical practitioner of the art at issue. I understand that courts consider a non-exhaustive list of factors, including (1) the educational level of the inventor; (2) type of problems encountered in the art; (3) prior art solutions to those problems; (4) rapidity with which innovations are made; (5) sophistication of the technology; and (6) educational level

of active workers in the field. I understand that it is not appropriate to consider the level of skill possessed by the inventor in defining the POSA.

29. I understand that whether a patent claim is obvious is ascertained at the relevant time period, which in this proceeding I understand to mean before January 26, 2012.

30. I understand that an obviousness analysis involves ascertaining the scope and content of the prior art, the background and experience of a POSA, the differences between the claimed invention and the prior art, and whether there are additional factors present that may nevertheless weigh against obviousness, such as unexpected results attributable to the invention actually claimed. I also understand that the obviousness analysis is based on the knowledge possessed by one of ordinary skill in the art at the time when the invention was made and that the use of hindsight is impermissible.

31. I understand that for a prior art reference, or a combination of references, to render a claimed invention obvious, the POSA must have had a motivation to modify or combine the references to arrive at the claimed invention. I also understand that separate from having a motivation to modify or combine, the POSA must have also had a reasonable expectation of success in modifying or combining the references to arrive at the claimed invention.

32. I understand that the following additional factors are probative—and at times the most probative evidence—in assessing obviousness: (1) the prior art taught away from the claimed methods; (2) the art expressed skepticism towards the claimed inventions; (3) a long-felt, unmet need existed that was resolved by the claimed inventions; (4) the industry praised the claimed inventions; (5) others copied the claimed inventions; and (6) using the claimed inventions yielded results that would have been unexpected based on the closest prior art. I understand that these factors are called objective indicia of non-obviousness or secondary considerations. I understand that a prior art reference teaches away from the claimed invention when a POSA, upon reading the reference, would be led in a direction that diverges from the one taken by the applicant. I understand that the skepticism factor considers whether industry or a POSA expressed skepticism towards whether or how a problem could be solved. I also understand that skepticism may be shown where industry or a POSA express skepticism that the claimed invention would work. I understand that evidence demonstrating a long felt but unresolved need supports the non-obviousness of a claimed invention because it is reasonable to infer that the problem would not have persisted if it were obvious to address. I understand that industry praise of the claimed invention reflects that a POSA would not have reasonably expected the claimed invention to succeed. I understand that the copying of the claimed inventions by competitors reflects their inventive and valuable nature.

I understand that obtaining results using the claimed inventions that are unexpected based on what was known in the prior art suggests that the claimed inventions are nonobvious.

33. I understand that there must be a nexus or connection between the claimed invention and the objective indicia of non-obviousness. I understand that nexus is presumed where the objective evidence is tied to a specific product that embodies the claimed features and is coextensive with them. I also understand that nexus may be shown where the objective evidence is the direct result of the unique characteristics of the claimed invention.

34. I understand that for a dependent claim to be anticipated or obvious, the independent claim from which that claim properly depends must be anticipated or obvious as well.

IV. TECHNICAL BACKGROUND

A. Genetic Information

35. All living organisms possess a genetic code, a chemical information system that contains the biological instructions required for an organism's development and function. Because an organism's genetic code dictates its biological functions, such as which proteins are produced, when they are produced, and where they are produced, understanding the composition and regulation of an

organism's genetic information is essential to understanding biological function generally.

36. An organism's genetic information is contained in deoxyribonucleic acids ("DNA") and ribonucleic acids ("RNA"). These nucleic acids are made of a chemically defined subunit termed a "base," which defines the identity of the nucleic acid. For example, canonical DNA is composed of four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). By organizing these bases into discrete linear sequences, nucleic acids can serve as an information system wherein the composition and order of the bases reflect a set of biological instructions that can be read by other cellular machinery.

37. Discrete sets of encoded information (such as the structure of a given protein) are retained in discrete regions termed "genes," which are contained within a cell's DNA. The totality of a cell's DNA is referred to as a cell's "genome." DNA is organized in a helical structure composed of two intertwined strands. At a given location in the genome, each DNA strand is a "complement" of the opposite DNA strand to which it is associated, which allows for both strands of DNA to contain the same genetic information—the first sequence information contained in the "template" strand and the same sequence information mirrored in the complement strand. As I describe in further detail in subsequent sections, the complementary

nature of nucleotides such as DNA is an integral phenomenon harnessed by nucleotide sequencing technology.

38. DNA is a molecule that encodes an organism's genetic blueprint. DNA contains all of the information required to build and maintain an organism. Nearly all of the cells within a single organism include exactly the same DNA. When required for actionable use within the cell, the genetic information contained in DNA is converted into RNA and then disseminated throughout the cell.

39. Like DNA, RNA is composed of nucleotides containing discrete bases: three bases—adenine (A), cytosine (C), and guanine (G)—shared with DNA and an additional base, uracil (U), corresponding to the thymine (T) base observed in DNA.

40. The genetic information contained in genes is converted to messenger RNA (mRNA) by an enzyme known as RNA polymerase during a process called transcription. mRNA is designed to efficiently traffic information from the DNA to sub-regions of the cell for processing into proteins. The entirety of a cell or organism's mRNA, including the genetic information contained therein, is referred to as its "transcriptome."

41. Complementary DNA (cDNA) is a form of DNA that is synthesized from an mRNA template in a process called reverse transcription. cDNA is DNA that is complementary to the mRNA and thus contains only the exonic sequences of the gene—those sequences that code for protein. This makes cDNA especially

useful in genetic studies because it represents the expressed genes in a cell, tissue, or organism, excluding introns and regulatory sequences found in genomic DNA.

42. In addition to studying the typical sequences of nucleic acids, researchers are also interested in atypical sequences. Accidental alterations in an organism's genetic sequence, termed "mutations," are associated with and often drive numerous diseases. For example, single base changes in an organism's DNA, known as "single nucleotide polymorphisms" (SNPs), can result in major downstream changes in protein expression and function, and in some cases can be the main driver of disease, such as sickle cell anemia. In additional cases known as "complex diseases," multiple SNPs or similar mutations throughout the genome combine to drive or exacerbate a disease state.

B. Nucleic Acid Genotyping

43. Genetic variation within a population arises when changes in an organism's DNA sequence occur that are then passed down to their descendants. When a specific variant is detected at a considerable frequency in a population—typically above 1%—it is termed a "genetic polymorphism." In contrast, "mutations" usually refer to the initial rare or novel alterations that arise spontaneously and can be associated with genetic diseases, but may or may not be passed down to descendants or may arise in a specialized setting (such as a cancerous tumor). A SNP represents one of the most prevalent types of polymorphisms.

Though many SNPs are clinically benign, some can markedly affect gene expression and protein function, potentially driving or contributing to disease, as exemplified by the single-base mutation in the beta-globin gene (HBB) responsible for sickle cell anemia, which has been known since the 1950s to be a disease with a genetic origin and etiology.

44. To systematically identify and characterize these variations, researchers developed the process of “nucleic acid genotyping,” which involves examining a segment of DNA and comparing it to a reference sequence or to the same segment in another individual. In diploid organisms—those that inherit two sets of chromosomes, one from each parent—each locus is typically present in two copies.² Therefore, an individual’s diploid genotype at a specific SNP site can be homozygous reference [AA], heterozygous [AB], or homozygous non-reference [BB], where “A” denotes the reference (or “wild type”) allele and “B” denotes the non-reference (or “minor”) allele. By systematically determining whether an individual has [AA], [AB], or [BB] at many such loci, scientists can create a detailed map of genetic variation relevant to traits or diseases.

² The exceptions are uniparental inheritance of the mitochondria and Y chromosome, which are only passed on from the maternal or paternal parent and are considered haploid.

45. This locus-specific approach enabled researchers to organize and interpret genetic data on a large scale, thereby identifying regions of the genome that play pivotal roles in health and disease. Because SNPs and other polymorphisms can provide insight into disease susceptibility, severity, and treatment response, scientists and clinicians benefit from methods that can genotype many loci simultaneously. To meet this need, high-throughput platforms were developed, including Affymetrix's large-scale genotyping arrays which are discussed in the following section, which rapidly evaluate tens of thousands—or even millions—of SNP positions in a single experiment. By focusing on known sites of variation across the genome, these arrays generate substantial volumes of data that can be used to identify disease-associated alleles (via genome-wide association studies), pinpoint genetic markers for clinical diagnostics, and facilitate personalized medicine.

46. Some of the earliest efforts to characterize genomic information included genotyping through genome walking, DNA microarrays, and sequencing.

C. Library Preparation

47. Before I discuss how prior art genome walking, DNA microarrays, and sequencing technologies worked, it is important to consider how genetic samples were prepared for use with these techniques.

48. Generally, after sample collection, researchers isolate the sample to be analyzed and prepare a library of that sample for further analysis. The procedures

used to generate that library are referred to as “library preparation.” Library preparation methods can vary for a number of reasons, but there is a general workflow when analyzing samples through genome walking, DNA hybridization arrays, and sequencers. Historically, this general workflow has consisted of fragmentation, adapter ligation, and amplification.

49. Fragmentation refers to a specific procedure in library preparation where larger samples are broken into smaller fragments through either physical or enzymatic means. Examples of physical fragmentation include nebulization (using compressed nitrogen or air to force samples through small holes), sonication (breaking covalent bonds with ultrasonic waves), and hydrodynamic shearing (using an aqueous solution to force samples through constrictions). *See* EX1001.0018 (6:60-63). There are numerous reasons why a researcher would fragment the sample during library preparation, including that the instruments used for genetic analysis typically only work with nucleic acid fragments. For example, Illumina sequence reads are governed by cycle lengths of typically 50, 75, or 150 nucleotides, yielding paired reads of (2x50, 2x75, or 2x150 base pairs in length). Typical fragment length was therefore 300-500 base pairs in length since at most the first (and last) 50, 75, or 150 base pairs of the fragment are sequenced by the instrument. Another reason is that the fragmentation process can separate fragments of interest from other genetic material, such as with enzymatic fragmentation, which targets specific

restriction sites (e.g., in the use for genotype-by-sequencing assessing genetic polymorphism).

50. Adapter ligation refers to attachment of adapters to the nucleic acid sample or fragments. There are numerous purposes for adapters, including for extension, amplification, and annealing fragments to instruments for analysis (such as the Illumina P5/P7 sequences which permit flow-cell binding).

51. Amplification refers to the increase of samples or fragments. It can be linear or exponential. It can also be conducted through thermal cycling or isothermal.

52. One method for amplification is polymerase chain reaction or PCR. Conceived by Kary Mullis in 1983, PCR utilizes enzymes known as polymerases and short oligonucleotides known as primers to catalyze the replication of new nucleic acid strands from a provided nucleic acid template. EX2040.0263.³ PCR accomplishes this replication through a series of reactions where the complement of a nucleic acid strand is generated. *Id.* PCR proceeds through iterative temperature cycling—denaturation, annealing, and extension—that conceptually doubles the number of target molecules with each cycle. *Id.*, 0265-267. This cyclical approach allows for exponential amplification of the nucleic acid strand. *Id.*, .0263.

³ I cite to the native page numbers of all Exhibits that have native page numbers.

53. In PCR, two primers are designed to flank the nucleic acid template, with the nucleotide sequences of the primers determining the reaction's specificity. *Id.*, .0265-0267. The first primer binds to a target site either located on the nucleic acid template or as part of an appended adapter, a short synthetic oligonucleotide that can be attached to the nucleic acid template. *Id.*; EX1009.0004. This process is known as annealing. The polymerase then extends from the 3' end of the primer by sequentially incorporating deoxyribonucleotide triphosphates ("dNTPs," a type of nucleotide) that are complementary to a provided template strand. EX2040.0263. This process is known as extension, and generates a nucleotide fragment that contains the primer and is complementary to the strand from which it is synthesized. *Id.*, .0265-0267. The extension product (as well as the complementary sequence of the original fragment) can then be bound by the second primer, and the process repeats itself, generating a second nucleotide fragment identical to the original template strand. *Id.* Over many PCR cycles, often 25–35 cycles, an initial nucleotide fragment can be exponentially amplified. *Id.*, .0263. The resulting products will contain the original nucleotide fragments along with an exponential number of copies of each fragment. These copies are often referred to as "amplicons."

54. PCR provided researchers with the ability to amplify low amounts of genetic material, providing access to genetic material that occurs in low amounts and

replenishing genetic material that could then be used for further experiments and methods. EX1009.0004. As primer sequences are introduced into the nucleotide fragments as part of the PCR process, specific synthetic sequences can be incorporated into the primers such that the synthetic sequences are introduced into the population of PCR products. *Id.* This allows researchers to manipulate PCR libraries to prepare them for downstream processing, such as introducing primer binding sites for sequencing primers. *Id.*, .0004-05.

55. While PCR expanded the capabilities of the field, it was also considered to be error prone. EX2075.0002; EX2041.0321; EX2036.0008-09 (31:20-33:13). Any one of several factors—including primers, polymerases, reaction temperature, and amplification cycles—contributed to the generation of off-products and other artifacts. EX2042.0008-11. For example, despite the development of high-fidelity polymerases exhibiting lower error rates, hundreds of millions of extension events across 30 cycles inevitably introduce accidental nucleotide substitutions and single-nucleotide insertions and deletions, into PCR products. EX2075.0002. Additionally, phenomena such as sequence-content bias and polymerase slippage and premature dissociation can lead to under- or over-representation of select PCR products or erroneous products. EX2042.0001. These errors are exponentially multiplied as more cycles of PCR occur, resulting in populations of amplified products containing

many artificial sequence errors introduced as part of the PCR process. Many of these erroneous products are referred to as “artifacts.”

56. Researchers attempted to minimize the errors associated with PCR. For example, numerous methods of PCR were developed as part of “genome walking,” including methods that introduced multiple sets of PCR cycles, additional components (like biotin) that could be used to capture fragments, and specialized adapters and primers that could suppress PCR amplification. *See, e.g.,* EX1007.1087. In general, each method demonstrated an improvement on the prior methods, but as explained in the following section, even later generation of methods still suffered from target specificity issues and amplification artifacts.

D. Anchored PCR and Genome Walking

57. The invention of PCR gave researchers the ability to selectively target genetic regions and amplify them. EX2040.0263. As early as the 1990s, researchers coopted this facet of PCR to develop methods to explore unknown regions of the genome, a technique known as “anchored PCR” or “genome walking” (when applied to genomic DNA). EX1007.1087. Researchers designed primers to target two known nucleotide sequences that flanked an unknown genetic sequence. *Id.* Typically, one of the known sequences was a either previously characterized endogenous sequences (sequences naturally present in the genetic sample of interest) while the other was a synthetic nucleotide sequences introduced into the sample, for

example by an adapter that was attached to a nucleotide fragment containing the unknown genetic sequence. *Id.*, EX2044.1644. Using PCR, the unknown genetic sequence in between the two primer sites could be amplified for further characterization. *Id.* Similar methods existed for RNA, which utilized a technique known as rapid amplification of cDNA ends (“RACE,” which converts RNA to cDNA) to introduce known sequences at the end of the resulting molecule. EX2043.0526. These known sequences could then be targeted in conjunction with a second primer targeting an internal portion of the cDNA. *Id.*, .0527.

58. Due to their design, anchored PCR and genome walking suffered from serious issues that hampered their effectiveness. As PCR-based methods that used primers to target specific regions of interest in a larger genetic sample, anchored PCR and genome walking relied on the specificity of the primers to accurately amplify the region of interest. EX1007.1087; EX2047.4363. Primer specificity refers to the ability of a given primer to bind to its intended target nucleotide sequence, as opposed to a) binding non-specifically to an unintended nucleotide sequence or b) binding the targeted nucleotide sequence at multiple locations in the sample, as nucleotide sequences may be repeated in genetic samples at different locations. The importance of primer specificity is compounded when using a synthetic sequence, such as an adapter as a primer binding sequence, as many methods introduce the synthetic sequence non-discriminately into the sample. *Id.*

In these cases, only the primer targeting the second, endogenous nucleotide sequence can be expected to confer specificity. EX2047.4363.

59. Non-specific primer binding, referred to as “off-target binding” is a major problem in genome walking because it leads to the amplification of unwanted DNA fragments, obscuring the desired target sequence.⁴ *Id.* This happens because primers can anneal to multiple locations in the genome, not just the intended target site, and with exponential amplification in PCR, these non-specific products can quickly overwhelm the specific product. *Id.* This makes it difficult to isolate and identify the specific DNA sequence of interest. Early anchored PCR and genome walking methods suffered from serious issues with off-target binding. EX2043.0527; EX2047.4363. Off-target binding was particularly problematic when primers targeted repetitive sequence elements, which are portions of nucleotide sequences that occur repeatedly throughout a sample. EX2043.0526 . Additionally, primer binding and PCR generally suffered from issues with certain nucleotide-rich sequences (e.g., A-T or G-C rich), which further narrowed the applicability of target-specific PCR methods such as anchored PCR and genome walking. *Id.*

⁴ This phenomenon, when discussed in the context of PCR amplification, is also known as “non-specific” amplification.

60. To address the issue of primer specificity, researchers modified the methods to optimize the PCR process to and limit the generation of off-target products. EX2043.0527; EX2045.0149. By the late 1990s, several methods existed that were designed to minimize non-specific amplification, including (1) PCR suppression, where non-desired PCR products are “suppressed” and not subsequently amplified further; (2) vectorette PCR, which utilizes special adapters with mismatched sequences that blocks amplification where a fill-in reaction does not correct for the mismatch; (3) splinkerette PCR, which is similar to vectorette PCR but uses a hairpin loop structure on the adapter in an attempt to promote further on-target specificity; and (4) nested PCR, a method wherein a second round of PCR is conducted using primers designed to bind selectively to the product of the first round of PCR. EX2044.1644; EX2045.0149; EX2046.0039. Generally, these methods were designed to either a) limit the ability of off-target products generated in a first round of PCR to be further amplified or b) add further rounds of primer targeting using additional primers to improve the specificity of the resulting PCR product. *Id.* Additional methods, such as adaptor long-range PCR, were also developed to further refine primer specificity-driven PCR in an effort to limit off-target products. EX2045.0149.

61. Most of these early methods relied on a combination of a first primer which targets a universal synthetic primer binding site introduced into the sample, a

second primer which targets a portion of the sample itself, and some additional feature to control for amplification. EX2043.0528; EX2044.1644; EX2045.0149; EX2046.0039. While these early, pre-NGS methods may have used a single target-specific primer, they demonstrated why the field had moved beyond single target-specific primers by January 2012:

- In 1995, Siebert *et al.* observed that “unpredictably primed PCR” (a method of characterizing a genetic sample by randomly annealing primers targeting common repetitive sequence elements and conducting PCR) and vectorette PCR “generated multiple PCR products even after nested PCR was performed.” EX1007.1087. Siebert explained that “the presence of multiple PCR products would complicate further characterization without Southern blot hybridization or extensive cloning.” *Id.* In an attempt to resolve this issue, Siebert developed a method that coupled vectorette PCR with PCR suppression. This method relied on special, vectorette primers, that were ligated to both sides of a DNA fragment. The vectorette primers contained a suppression tag that would bend fragments into a hairpin loop, suppressing exponential amplification. I discuss this further below. *See infra* ¶¶ 200-204. Siebert determined that this method improved target specificity but that the specificity was “due to the suppression effect of the adaptor/adaptor primer combination.” EX1007.1088.

- Also in 1995, Devon *et al* published the splinkerette PCR method, designed to address issues with “[i]llegitimate products result[ing] from non-specific annealing of either primer” in vectorette PCR reactions. EX2044.1644. Devon explained that “[e]xtension across a vectorette sequence at either end of the insert molecule will result in the production of a sequence complementary to the vectorette primer, allowing exponential PCR amplification without involvement of the specific target DNA primer.” *Id.* To address these issues with vectorette PCR, Devon proposed “splinkerette PCR,” which incorporates a hairpin structure on the bottom strand of the adapter. *Id.* Devon explained that this more specialized adapter improved amplification efficiencies, especially for larger DNA fragments, due to the modified structure of the adapter. *Id.* I have included below a diagram from Devon that reflects the difference in structure for the vectorette and splinkerette adapters.

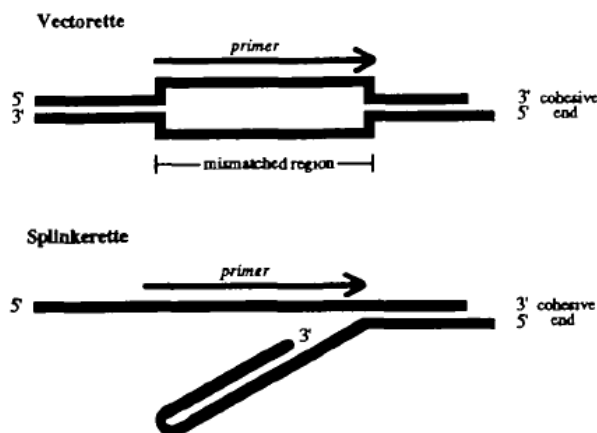


Figure 1. Schematic diagram of the vectorette and splinkerette.

Id.

- In 1996, Chenchik *et al.* (including Siebert) disclosed a method to a RACE protocol designed to selectively target the ends of cDNA molecules. EX2043.0526. The method introduced adapters at the end of the cDNA molecule that could be used in conjunction with a primer targeting an internal sequence to PCR the ends of the fragments. *Id.* The adapters were designed to only function in conjunction with the target-specific primer, and would suppress any fragment that contained two full-length adapter sequences. *Id.*, .0528. The method was further combined with a long and accurate PCR method. *Id.*, .0526. Chenchik observed that prior RACE methods suffered from “high background of nonspecific PCR products.” *Id.* In a proof of concept experiment, Chenchik observed a reduction, but not elimination, in nonspecific PCR products. *Id.*, .0529-531.

- In 1998, Padegimas *et al.* reported an adapter ligation-based walking technique designed to improve PCR specificity using an optimized adapter ligation strategy while removing (via degradation) un-ligated fragments. EX2045.0149. The adapters were designed with three discrete structural components to enhance stability and prevent degradation during the newly implemented degradation step, wherein enzymes were added to degrade non-ligated nucleic acid fragments in an effort to enhance PCR specificity. *Id.*, .0151, 153. Following adapter ligation and non-target degradation, primers would then be used to target the adapters and a target sequence contained within the ligated nucleic acid fragments for PCR. *Id.*, .0151. Padegimas states that the method was developed to address “low specificity, low efficiency, and/or extensive cycling that contributed to a higher error rate in PCR products” in current PCR based walking methods, including vectorette, splinkerette, and adapter-based methods. *Id.*, .0149. Padegimas reported a reduction in off-target products with its method, although the method required high degree of optimization to achieve this reduction. *Id.*, .0152.
- In 2005, Miura *et al.* reported that “PCR applications [that] use only a single unique primer in combination with a common primer [] includ[ing] rapid amplification of cDNA ends (RACE) [], PCR-mediated genome walking [], adaptor-tagged competitive PCR (ATAC-PCR) [], and so forth” were “often

plagued with non-specific amplification products derived from so-called ‘off-targets.’” EX2047.4363. To address this, Miura sought to develop an algorithm to determine primer binding specificity performance. *Id.* In two proof-of-concept experiments, Miura observed that off-target priming could be reduced when primer design was optimized, but not eliminated. *Id.*, .4366-68. Further, Miura observed that off-target primer binding was sometimes exacerbated using the algorithmic design. *Id.*, .4369.

- Also in 2005, Tsoktouridis *et al.* compared various PCR methods and found “no simplicity, low reproducibility, high backgrounds of nonspecific PCR products, insufficient homopolymer tailing, small PCR product amplification, and laborious time-consuming procedures.” EX2048.0885. As Tsoktouridis explained, techniques “using adaptors for PCR amplification of unknown DNA fragments flanking a known sequence” (genome walking methods) were “limited by a relative short walking distance, low specificity, low efficiency, and/or extensive cycling, which contributed to a higher error rate in PCR products.” *Id.*, .0886. To circumvent these issues, Tsoktouridis developed an optimized adaptor long-range PCR method, wherein adapters were designed for use with specific digestion enzymes (enzymes designed to cut DNA at certain locations) after which long-range PCR could be conducted using a primer that targeted the adapters. *Id.*, .0886-87. Additionally, Tsoktouridis

observed that further optimization of the amount of sample used with the methods was required to reduce off-target primer binding. *Id.*, .0887.

- In 2011, researchers recognized that despite that “many studies have been made to improve the genome walking method,” “it is still in need of improvement.” EX2050.0002. Sohn *et al.*, commenting on the field in 2011, explained that the “most serious problems of the PCR-based genome walking methods are low amplification efficiency and non-specific amplification.” *Id.* Sohn noted that to “improve this problem, complicated multi-step methods were developed, including secondary PCR, purification of desired PCR products using biotinylated primers, or improvement of PCR cycle conditions” however, even then, “satisfactory performance [had] not yet been achieved.” *Id.* Sohn proposed yet another modification to the traditional genome walking method, utilizing a nucleotide analog ligation approach coupled with PCR, wherein the ends of the nucleic acid fragments in a sample were chemically modified to introduce a nucleotide chemical analog that could be target to enhance adapter ligation. *Id.*, .0003. After which the adapter could be used in conjunction with a target specific primer to target via PCR a region of interest within the fragment. *Id.*

62. Indeed, as demonstrated above, despite over sixteen years of method development, primer specificity remained a serious issue limiting the utility of

primer-specific PCR methods. Over time, the art moved away from using single target-specific primers without many additional safeguards and towards specialized methods, such as complex PCR suppression schemes. EX2048.0885-86. By the mid 2000s, algorithmic design approaches were also being developed to optimize primer design in an effort to reduce primer off-target binding. EX2047.4363. However, even these efforts failed to resolve off-target issues, and it was generally accepted that primer specificity issues had to be mitigated by increasingly complex PCR assays and optimized primer design. EX2048.0885-86. As such, off-target primer binding remained an unresolved issue inherent in PCR applications. Indeed, by the late 2000s, researchers were considering alternatives to PCR amplification to circumvent these issues and enable multiplex sequencing of targeted gene regions.

E. PCR-Free Sequencing Methods

63. As discussed above, despite the tremendous utility of PCR, issues with logistics, sequence errors, primer specificity, and amplification bias limited its application. EX2042.0001-02; EX2051.0050, 0052; EX2052.0001-02. These issues were particularly pronounced in applications assessing the composition of a genetic sequence, as these PCR issues could convolute or artificially bias the resulting sequence analysis. EX2042.0001-02; EX2052.0001-02. With the advent of next-generation sequencing methods (discussed in further detail below), PCR issues became even more pronounced, as many NGS methods utilized PCR for both sample

preparation and as part of the sequencing process. *Id.* Further, as the scale of sequencing methods increased, inherent limitations in PCR applications became apparent as primer-specific approaches were not considered scalable to whole genome assessments. EX2051.0050, 0052; EX2052.0001-02; EX2066.0182.

64. Given these issues, many researchers began to consider alternative approaches for handling genetic material. *Id.* Several proposed incorporating required sequencing components directly into adapters in order to eliminate the need to introduce them during PCR (allowing researchers to forego PCR altogether). EX1009.0019; EX2053.1005-08; EX2054.0291-92. Methods were also developed to minimize the amount of PCR amplification required prior to sequencing by replacing certain PCR steps with selective capture methods (molecular techniques designed to capture sequences of interest in a larger pool of sequences for downstream assessment). EX2052.0001-02. Still other researchers sought novel hybrid array methods without any PCR amplification step. *Id.*, .0011. By 2011 PCR amplification was widely recognized as the predominant source of error in standard sequencing workflows. EX2042.0001-02. Indeed, issues stemming from PCR amplification in NGS sequencing were such that companies developed commercial sample preparation kits designed to limit PCR amplification during sample preparation for NGS applications. EX1008.0241; EX2056.0003.

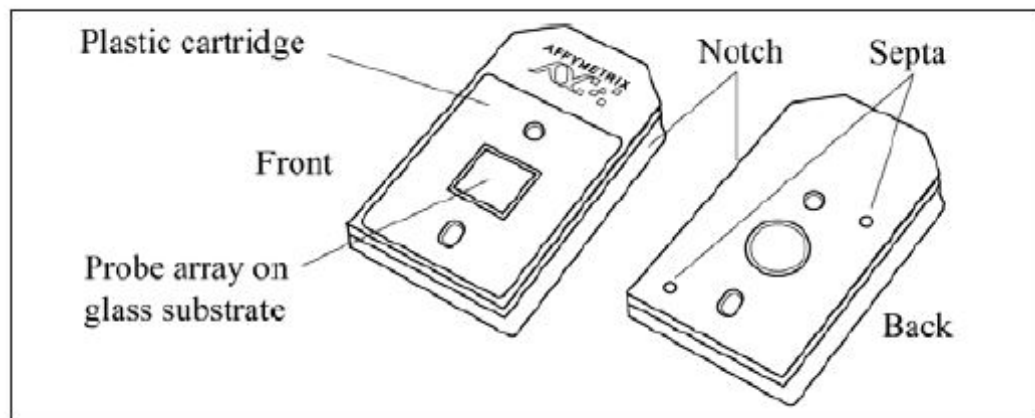
F. Hybridization Arrays

65. One key high-throughput genotyping approach involved hybridization arrays,⁵ which exploit the principle that complementary nucleic acid strands selectively bind to one another (e.g., adenine pairing with thymine, guanine pairing with cytosine). In the early 2000s, Affymetrix employed a proprietary photolithographic manufacturing process—inspired by semiconductor fabrication—to synthesize short 25-mer oligonucleotide probes directly on a silica or glass substrate. Through iterative cycles of UV light exposure through custom-designed photomasks, nucleotides were selectively deprotected and added in a precise, stepwise manner, creating microscopic “features,” each containing up to a million copies of a single probe sequence. On arrays designed to genotype more than 10,000 SNPs (such as the Mapping 10K array), each SNP was interrogated by over 40 probes, incorporating perfect-match and mismatch designs. Consequently, a single array featured over 400,000 distinct probe sequences, which—multiplied by the millions of replicated copies per feature—resulted in hundreds of billions of physical probe molecules on one chip. During a genotyping assay, only DNA fragments complementary to these probes would hybridize and remain on the array after washing, enabling accurate, parallel determination of whether an individual was

⁵ Hybridization arrays are also referred to as microarrays or DNA density arrays.

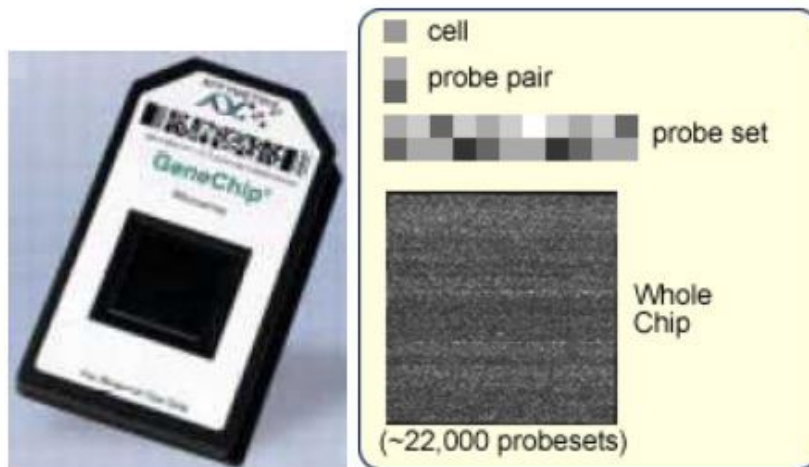
homozygous reference [AA], heterozygous [AB], or homozygous non-reference [BB] at each interrogated locus. Affymetrix thus enabled high-throughput, highly specific genotyping (~99.8% accuracy) across tens to hundreds of thousands of highly polymorphic SNPs.

66. These arrays were designed for high-throughput screening of known SNPs and could assess hundreds of thousands of SNPs in a single assay. For example, the Affymetrix Genome-Wide Human SNP Array 6.0 (depicted below) consisted of a single array that enabled the genotyping of approximately 900,000 SNPs and contained over 6.8 million unique probes, each replicated hundreds of thousands to millions of times on the array surface.

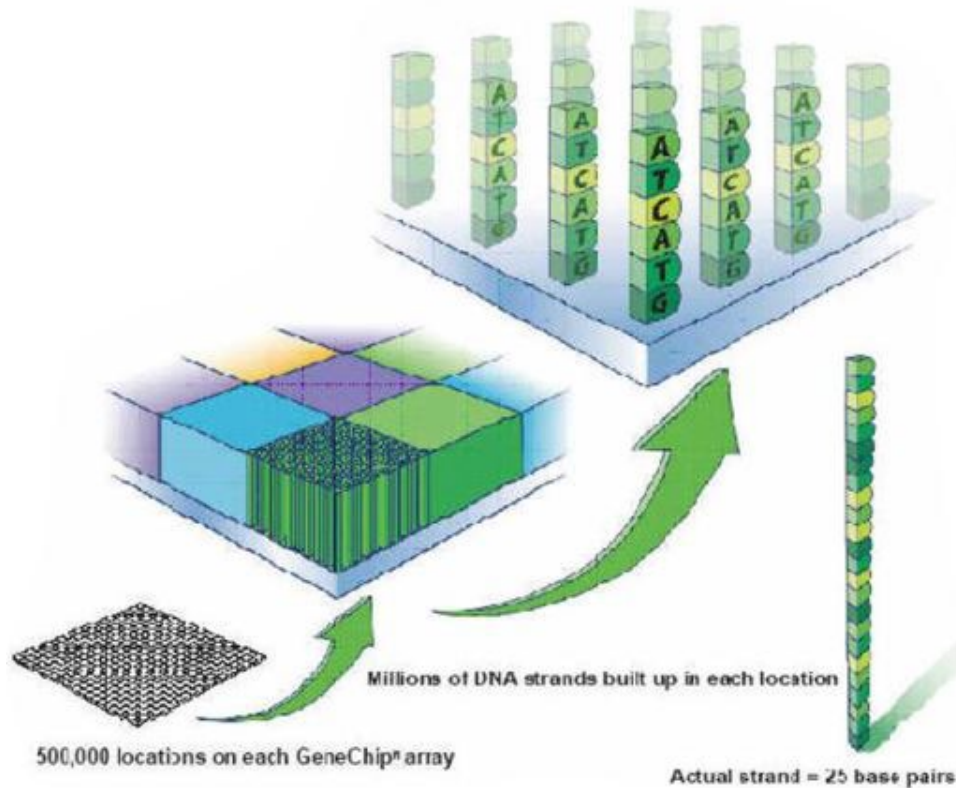


EX2001.0002.

67. Affymetrix produced a series of hybridization arrays known as GeneChips, designed for large-scale genetic analysis. An exemplary GeneChip is shown below, along with a depiction of the single array:



EX2002.0001. Affymetrix continually increased the probe density, incorporating hundreds of thousands to millions of different probes (unique oligonucleotide sequences) onto a single chip, with each 25-mer replicated many times, resulting in hundreds of billions of probes in total. A schematic depiction of the array is shown below:



EX2003.0001. Because each probe was synthesized at a known, fixed location on the array, researchers could interpret hybridization patterns using specialized software to identify genes, relative expression levels (for expression arrays), or genotypes at specific SNPs (for genotyping arrays).

68. To use a GeneChip, a researcher first needed to prepare the sample, often through DNA fragmentation, amplification, and labeling. The prepared sample would then be injected into the array through septa on the back of the device.

EX2004.0006. Once loaded, the sample was allowed to hybridize within the enclosed chamber. The workflow to genotype a sample via the GeneChip Platform is depicted below.

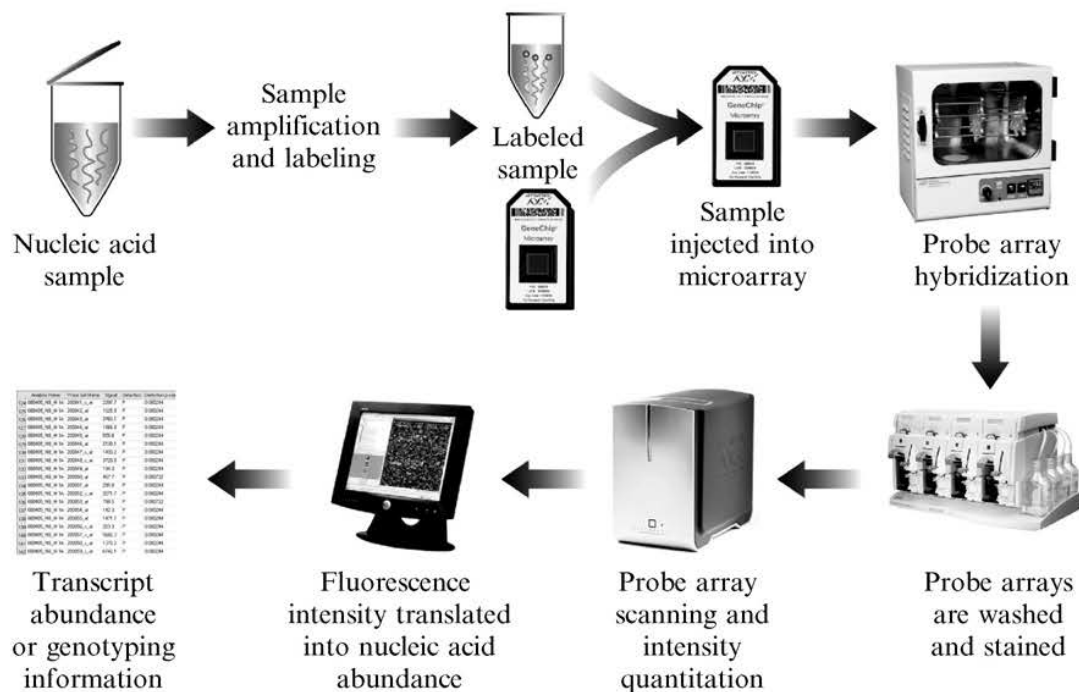


FIG. 1. Flowchart of a GeneChip System microarray experiment. Once the nucleic acid sample has been obtained, target amplification and labeling result in a labeled sample. The labeled sample is then injected into the probe array and allowed to hybridize overnight in the hybridization oven. Probe array washing and staining occur on the fluidics station, which can handle four probe arrays simultaneously. The probe array is then ready to be scanned in the Affymetrix GeneChip scanner, where the fluorescence intensity of each feature is read. Data output includes an intensity measurement for each transcript or the detailed sequence or genotyping (SNP) information.

Id.

69. As part of sample preparation, the DNA to be assessed by the array was typically fragmented using restriction enzymes selected for each particular Affymetrix array design (e.g., NspI, StyI, HindIII, or XbaI). For instance, certain early arrays such as the 10K array relied on XbaI or HindIII, while the Genome-Wide Human SNP Array 6.0 employed both NspI and StyI to fragment genomic DNA near the SNPs of interest. After fragmentation, adapters containing common primer-binding sites were ligated to the DNA fragments, enabling PCR-based

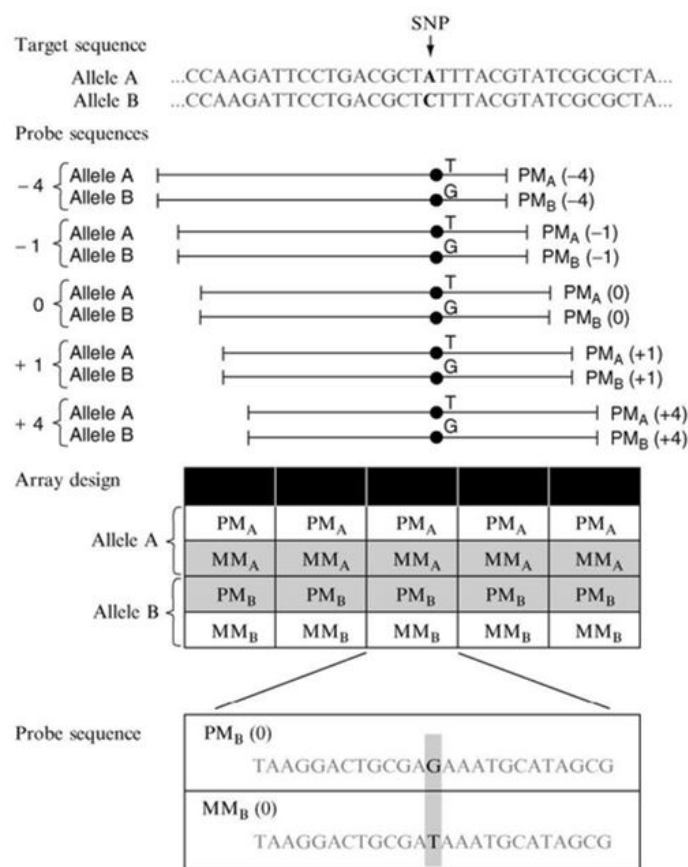
amplification to generate sufficient material for genotyping (EX2004.0015–16). This amplification was size-selective, enriching for fragments within a defined range and thereby standardizing the final library. Such size standardization helped ensure more consistent and accurate genotyping results across the entire genome.

70. After amplification, the DNA was denatured to produce single-stranded fragments. These fragments were then washed over the silica chip, allowing each fragment to bind (hybridize) to complementary oligonucleotide probes. Specifically, probes were designed to discriminate between possible alleles at a SNP locus. For instance, if a SNP of interest was heterozygous [AB], containing both the reference (A) and minor (B) alleles, fragments bearing allele A would hybridize to probes specific for allele A, while fragments bearing allele B would bind to probes targeting allele B. This allele discrimination process underpinned Shapero's use of hybridization arrays.

71. In many cases, probes targeting a specific allele were replicated (duplicates or multiple similar sequences) to increase the aggregate signal. Additionally, slight positional shifts (e.g., -4, -1, 0, +1, +4) were built into the probe design, ensuring robust detection of the SNP by accounting for neighboring sequence context. Often, arrays also employed mismatch probes (MM), which incorporated a deliberate non-complementary base at the SNP locus. These MM

probes served as controls to account for non-specific binding and background signal, thereby refining the accuracy of genotype calls.

72. A representative figure demonstrating array design, including both mismatch probes (denoted as “MM”) and probes designed to vary slightly in target sequence location (denoted as “-4,” “-1,” “+1,” and “+4;” indicating a shift in location from the reference probe denoted as “0”), is provided below:



EX2004.0012. These carefully arranged probes collectively produced a detailed hybridization signature that algorithms could interpret to assign genotypes.

73. Once the prepared DNA had hybridized to the array, each bound probe produced a fluorescent signal corresponding to the presence of complementary DNA. In practice, a biotin label was introduced into the DNA near the end of the amplification process. After hybridization, a streptavidin-phycoerythrin reagent was applied, which bound to the biotin-labeled fragments and generated fluorescence at the probe's location. A scanner or computer-driven imaging system then quantified signal intensities for each probe, enabling the user to analyze the entire microarray and to determine the alleles present at each locus (EX2004.0006). Complex algorithms were developed to optimize signal-to-noise ratios, correct for background fluorescence, and reliably call [AA], [AB], or [BB] genotypes based on the intensity patterns.

74. While genotyping by hybridization arrays provided a powerful, high-throughput platform for genetic assessment, the arrays were inherently targeted in nature. Probes were designed with high specificity to known target sequences, so the arrays could only query the particular SNPs for which they had been manufactured. Additionally, in order to design SNP probes, the location and flanking sequence of the SNP had to be known in advance. This requirement limited the applicability of microarrays in exploratory genomics where the relevant sequences or variants were unknown, driving the development of next-generation sequencing approaches that permit de novo discovery of genomic alterations.

G. Nucleic Acid Sequencing

1. Pre-NGS Technologies for Sequencing

75. Given the role of the genetic code in governing biological function, developing methods of identifying genetic sequences in various biological contexts has been a primary goal of scientific research. Early methods focused on DNA sequencing and technical methods for “reading out” the nucleotide composition of individual DNA strands. Early sequencing methods were time- and resource-consuming and were not suited to scalable applications. These limitations made sequencing large polynucleotides a technical challenge and genomes, basically infeasible. These issues were somewhat ameliorated by the advent of large scale-cloning and/or the parallel sequencing of many regions of DNA that would then be assembled together into a “working draft” of the genome.

76. One of the first sequencing methods, “Sanger Sequencing,” also known as the chain termination method, was developed by Frederick Sanger in 1977. Before the advent of Polymerase Chain Reaction (“PCR”) technology, Sanger Sequencing was performed directly on purified DNA templates, typically from cloned DNA fragments. Sanger Sequencing involved the enzymatic synthesis of DNA strands and their controlled termination using dideoxynucleotides (“ddNTPs”).

77. There were major challenges and limitations to this approach. The need to clone and grow bacterial cultures to obtain sufficient DNA for sequencing made

the process cumbersome and slow. The quality and length of the sequence reads were limited by the efficiency of the cloning process and the resolution of the electrophoresis gel. The method was also less sensitive and less scalable compared to later techniques that utilized PCR for amplification, making it challenging to sequence very small amounts of DNA or conduct large-scale sequencing projects efficiently.

78. The advent of PCR revolutionized Sanger Sequencing. Through PCR, researchers were able to take nucleic acid samples and amplify those samples without the need to clone and grow bacterial cultures. For DNA samples, PCR consists of three main “phases,” dictated in part by the temperature of the reaction mixture: denaturing, annealing, and extending. The denaturation phase separates the double-stranded DNA into single strands. The annealing phase allows for primers to bind to the single-stranded DNA to direct polymerase loading and facilitate the start of the nascent strand. The extension phase allows the polymerase to extend the nascent strand from the primer through the addition of nucleotides complementary to those on the template strand. Completion of all three steps constitutes one PCR “cycle,” after which a copy of the template strand has been produced. By running a series of PCR cycles, the original DNA sample can be exponentially amplified, as the strands produced in each PCR cycle can then serve as template strands in the following cycle. *See* EX2005.0344.

79. Another innovation to Sanger Sequencing was its automation through platforms such as the ABI 3730 DNA Analyzer, which transformed the landscape of genetic analysis by integrating advanced automation and fluorescence-based detection technologies. Sequencing samples were prepared in a mix that included the purified DNA template, a sequencing primer, DNA polymerase, and a mixture of regular dNTPs alongside specially dye-labeled dideoxynucleotides (“ddNTPs”). Each type of ddNTP, which terminates DNA chain elongation upon incorporation, was tagged with a distinct fluorescent dye. This labeling allowed for the simultaneous identification of each nucleotide during the sequencing process, despite the reactions occurring in a single tube. The data collected from the ABI 3730 was processed using software that translated the signals into a readable sequence displayed as a chromatogram, which could be interpreted to identify the DNA sequence. This automated system reduced the manual effort involved in sequencing and increased both throughput and accuracy.

80. Sanger Sequencing was limited to sequencing a single DNA molecule at a time in a sequencing reaction. Even with the advent of automated Sanger Sequencing, this structural aspect of Sanger Sequencing limited its sequencing throughput, making sequencing at scale both costly and logistically challenging. For example, completion of the Human Genome Project took approximately 13 years and \$2.7 billion, while the National Human Genome Research Institute estimated

that sequencing a second human genome using the Sanger Sequencing methods available in 2003 would cost another \$50 million. *See* EX2006.0001, 06. These technical limitations rendered general application of Sanger Sequencing to large or complex genomic sequencing tasks extremely expensive and time consuming, limiting genome-scale activity (in the US at least) to a dozen or so nationally-designated sequencing centers funded through large infrastructure grants from NIH, DOE, and the like.

2. Next Generation Sequencing

81. Coinciding with the completion of the Human Genome Project in the early 2000s, great scientific and public interest in a deeper understanding of the genome led to the commercial development of more affordable and scalable techniques for genomic assessment. These efforts resulted in a series of novel technologies and methods collectively termed “Next-Generation Sequencing (‘NGS’) Technologies.” EX2007.0333. As compared to traditional sequencing methods such as Sanger Sequencing, NGS technologies were characterized as cheaper, more efficient, and capable of assessing both larger genetic samples and longer portions of the genome in high fidelity. As compared to prior methods, NGS sequencing was defined in part by its “massively parallel” nature—i.e. its ability to process tens of thousands to millions of DNA sequences at the same time. Most NGS technologies utilized a “sequencing-by-synthesis (SBS)” strategy.

EX2008.0213-14. SBS leveraged the sequential nature of genomic information to assess each nucleotide in a sequence in order. Polymerases were coupled to a signaling output that allowed for the identification of the incorporated nucleotide.

82. Compared to traditional sequencing methods, early NGS technology afforded researchers several benefits. Due to their design, NGS systems allowed researchers to sequence thousands of reads in tandem in an automated manner, increasing both the platform capacity and the depth and coverage of the generated sequencing data. This made genetic sequencing both cheaper and more scalable as compared to prior methods. The parallel sequencing paradigm allowed researchers to sequence multiple genes and assess millions of variants simultaneously, while reducing the required DNA input and, due to the scaled approach, decreasing sequencing times. These factors led to an overwhelming increase in the amount of sequencing data generated in the course of research and a deeper understanding of the structure and content of many genomes across species. EX2007.0333.

83. Pyrosequencers, such as the 454 sequencers, were among some of the first set of platforms that harnessed this “second generation” sequencing technology. Illumina produced the most commercially successful sequencing instruments in this class, which also included Ion Torrent, Helicos, and ABI/Lifetech SOLiD (among others). Over time, NGS sequencers improved on their ability to process longer read

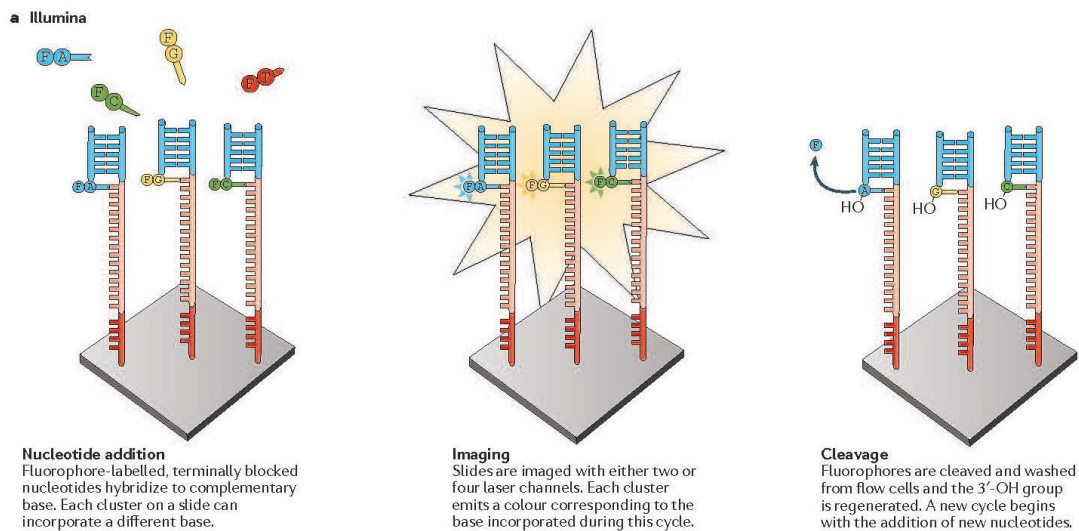
lengths, which resulted in a “third generation” of sequencers, which included Nanopore and PacBio sequencers.

84. Around January 2012, 454 sequencers, Illumina, and SOLiD were considered to be the three platforms that were “widely available through academic core laboratories and commercial service providers.” EX2057.0003. These three platforms “traditionally split their focus into fewer long reads (454) vs. more short reads (Illumina and SOLiD).” *Id.* A POSA would have understood that long reads were “optimal for initial genome and transcriptome characterization because longer pieces assemble more efficiently than shorter pieces” but that shorter reads would have been favored for more targeted sequencing, especially given the diminished accuracy of sequencing longer reads at the time. *Id.*

85. Even before January 2012, Illumina dominated the market. *See* EX2060.4328 (“As of this writing, Illumina dominates the market with 60% of the second-generation sequencer installations, while Life Technologies’ SOLiD system and Roche split nearly all of the remaining market at 19% each.”). Illumina sequencers employed a “cyclic reversible termination” sequencing method. EX2007.0335. Samples to be sequenced were first prepared to construct a sequencing library. As with the 454 sequencer, library preparation varied but generally involved the fragmentation of genetic material followed by adapter ligation and PCR to introduce Illumina platform-specific adapters into the sample.

Using these adapters, sequencing library was then attached to a flat, solid surface on the Illumina sequencing instrument via hybridization of the adapters to short complementary sequences located on the sequencer surface. *Id.*, .0336. The attached sequencing library was then amplified on the surface of the sequence to generate copies of the sequencing library. *Id.* Following cluster generation, sequencing commenced. Sequencing began by priming the nucleotide sequencings using universal Illumina primers so that nucleotides could be incorporated into a recently generated complementary stand. *Id.*, .0335. Sequencing occurred in a series of cycles. *Id.* Unlike single-nucleotide addition sequencing approaches, all four nucleotides were provide for each cycle such that any of the four could be incorporated into the elongating complementary strands. *Id.*, .0335-36. Each of the four nucleotide species was labeled with a different fluorescent label. *Id.* Additionally, all of the nucleotides were chemically modified such that no nucleotide could be added after incorporation of a previous chemically modified nucleotide. *Id.* As such, in each cycle only one of the four chemically modified nucleotides were incorporated into each elongating strand. Residual unbound nucleotides were then washed away and the fluorescent signal of each chemically modified nucleotide (reflecting the identity of the nucleotide) was recorded. *Id.* The fluorescent label and chemical modification were then removed, allowing for another nucleotide to be added to the nascent complementary strands, and the cycle was repeated. *Id.*

Using this process, the identify of each nucleotide incorporated into the complementary strands of the sequencing library could be recorded, and the sequence of the sequencing library could be determined. A figure representing the Illumina sequencing process is provided below.



EX2007.0340.

86. While these second generation NGS platforms were viewed as an overall improvement to sequencing technology, these platforms still failed to adequately address the concerns in the art at the time. For example, these NGS platforms exhibited a higher error rate than Sanger Sequencing per read. EX2007.0333. Additionally, compared to traditional methods, many NGS methods generated significantly shorter individual sequencing reads, while long-read sequencing approaches remained expensive and resource intensive. *Id.* In addition to these limitations, the sheer volume of genetic material that could now be

sequenced exacerbated issues with being able to isolate specific regions of genetic material for analysis. EX2008.0215-16. To resolve this issue (among other issues), methods were developed to narrow the window of genomic assessment, such as through “targeted enrichment.”

H. Target Enrichment

87. The inherent size and complexity of the genome have presented longstanding issues when attempting genetic sequencing. Often, researchers and clinicians are interested in select regions of the genome, such as those known to be associated with particular diseases or looking to compare a tumor DNA to normal DNA to detect changes in specific genes, that are minute in size compared to the genome as a whole. Sequencing methods can struggle to sequence smaller regions of interest in an efficient manner while ensuring sufficient signal (region of interest) to noise (signal from other genomic regions) ratios in the sequencing data and minimizing off-target effects (reactions at other regions of the genome outside the area of interest), all of which are necessary to generate accurate sequencing data. Researchers recognized that NGS “possesse[d] a fundamental weakness—the inability to easily target specific regions of a genome.” EX2009.0001.

88. In order to minimize these issues, NGS pipelines were integrated with targeted enrichment approaches designed to enrich a genomic region of interest relative to the rest of the genome prior to sequencing. As understood in the art, the

term “enrichment” means to selectively increase the proportion of one part of a sample or reaction mixture over other portions of the sample or reaction mixture. *See, e.g.*, EX2010; EX2011 (describing the use of various prior art target enrichment methods to enrich for target regions). Targeted enrichment methods sought to amplify the nucleotides corresponding to the region of interest while removing or minimizing those reflecting unwanted regions of the genome, thereby enriching the amount corresponding to the signal of the region of interest relative to the rest of the genomic sample.

89. As of the priority date of the '012 and '108 patents, three major categories of target enrichment methods existed: capture-by-hybridization, polymerase chain reaction (PCR)-based methods, and capture-by-circularization. *See* EX1001.0016 (1:39-47). Each is discussed in further detail below.

1. PCR-based Target Enrichment Methods

90. PCR-based target enrichment methods, as the name suggests, employed PCR to enrich regions of interest from a nucleotide sample prior to sequencing. *See* EX1001.0016 (1:46-50). Primer pairs were designed to target a specific region in a nucleotide sample such that the region was amplified by PCR, while nucleotide sequences outside of the region defined by the primer pair were not amplified. The method could be scaled combining groups of primer pairs that targeted discrete regions to amplify multiple target regions simultaneously, a method termed

“multiplex PCR.” *Id.* Multiplex PCR was useful for targeting small-to-medium-sized genomic regions. PCR-based enrichment was often quicker and more cost-effective than hybridization-based capture approaches but was generally suitable for smaller genomic regions due to limitations in multiplexing capacity and primer design. *See* EX1001.0016 (1:51-53).

91. PCR approaches required that sequences within or near the region of interest were known such that primer pairs could be designed to target the region, often precluding true exploratory studies. *See* EX1001.0016 (1:46-50). Additionally, primers were often not entirely specific to the primer target sequence (the exact complementary sequence of the primer) and could bind to genomic or complimentary DNA with similar, but not exact, sequence homology. *See* EX1001.0016 (1:51-53). When sequencing regions of a genome, this issue led to off-target binding to similar sequences outside the region of interest, both generating confounding sequence products that made sequencing interpretation difficult (or impossible) and sequestering primers that would otherwise bind to the actual target sequence. Further, because both the genome and transcriptome are characterized by frequent repetitive sequences, primers designed to target a sequence in a region of interest could inadvertently bind to the same sequence elsewhere in the genome, leading to further off-target binding. EX2012.0037. As such, primer-based enrichment methods were often characterized by low specificity and high

sequencing background noise. These issues were more pronounced with NGS where the platforms could not target or filter samples. *See, e.g.*, EX2013.1799 (“All PCR-based NGS systems share the common disadvantage of bias introduced during amplification. . . . bias [is] introduced during amplicon library preparation.”).

92. QIAGEN, which designed multiplex PCR kits at the time, published prior art handbooks discussing these concerns. For example, in the October 2010 QIAGEN Multiplex PCR Handbook, QIAGEN explained that “PCR is an exquisitely sensitive technique that can be used to amplify and detect even a single copy of a nucleic acid sequence,” but that “amplification of such a low number of target sequences is often limited by the generation of nonspecific PCR products and primer-dimers.” EX2014.0037.

93. In addition to the specificity issues discussed above, PCR-based target enrichment methods suffered from design issues that limited their scalability. *See* EX1001.0016 (1:51-53). PCR-based target enrichment methods employed highly parallel PCR amplification, which required that each target sequence corresponded to a pair of unique, sequence-specific primers. *See* EX1001.0016 (1:48-53). At the time, researchers recognized that “primer design is a critical step in all kinds of RT-PCR methods to guarantee specificity and efficiency of a target amplicon.” EX2015.0001.

94. The prior art at the time taught that it was necessary to use primer pairs rather than a single target specific primer. For example, in 1999, Matz reported issues with “amplification produced by a single primer” and proposed PCR suppression methods to minimize such amplification. EX2016.1558. As another example, in 2008, Bru reported that a single primer mismatch could “greatly decrease PCR efficiency” and taught that “the use of multiple sets of primers targeting different subgroups may be preferable when the sequence polymorphism of the targeted gene family is high.” EX2017.1663. Unfavorable sequencing outcomes using single target-specific primers led researchers to outright reject their use, including one research group that wrote:

While the specificity of individual primers may be testing using BLAST, no further consideration may be given to the influence of non-target sites that can result in suboptimal binding of the primers. ***Accurate quantification requires primer sets that facilitate maximum amplification efficiency.*** In our experience, it is usually necessary to design, synthesise, and validate several primer pairs, until a set is obtained that generates no primer dimers and results in near 100% amplification efficiency.

EX2018.0598 (emphasis added). Other researchers turned to other techniques entirely, such as hybrid capture, to avoid the use of PCR primer pairs. *See, e.g.*, EX2019.0329 (“Unlike PCR-based enrichment methods where both the 3' and 5' ends of targeted DNA must be specified through the use of template-specific primers, hybrid capture allows for the enrichment of sequences containing only partial homology to the viral insertion site.”). Before the claimed inventions, the widely

held view in the field was that primer pairs were required to facilitate effective enrichment through amplification, especially PCR amplification. In fact, I am aware of no instances before the priority date in which one target-specific primer was used in connection with PCR amplification to enrich for target sequences.

95. While possible to pool primer pairs together, multiplexing faced several issues. For example, different primer pairs sometimes required different amplification parameters (i.e., temperature and duration settings that facilitate efficient amplification). *See* EX1001.0016 (1:51-53). These diverse amplification parameters were difficult, if not impossible, to uniformly account for in a given PCR as the amplification conditions were often set and held consistent throughout the PCR cycles. Additionally, pooling multiple primer pairs in the same PCR reaction compounded the issues with primer specificity, as previously discussed. Similarly, pooling multiple samples could lower primer sensitivity on a per-sample basis, as multiple samples competed for primer pairs. Moreover, the introduction of more and varied primers exacerbated the issue of primer dimers, a phenomenon wherein primers bind and amplify each other rather than the true primer target sequence, generating unwanted PCR byproducts and sequestering primers, thereby lowering amplification efficiency. EX2020.3235.

96. In part to address these limitations of multiplex target enrichment by PCR, researchers developed an enrichment method known as “microdroplet-based

PCR enrichment” (“microdroplet PCR”) which could be performed using microfluidic devices. *See* EX1001.0016 (1:53-58). Microdroplet PCR required the generation of small solution droplets that contained discrete primer pairs designed for the genomic regions of interest. *Id.* Fragmented DNA samples and PCR reaction mix components could then be introduced into the droplets via a microfluidic chip, which maintained the individual integrity of the primer droplets, generating thousands of discrete PCR reactions while maintaining primer-pair separation. Following targeted enrichment, the droplets could be broken to mix the enriched products into a general solution for sequencing.

97. Microdroplet PCR demonstrated comparable accuracy and coverage to singleplex PCR sequencing while limiting issues with primer specificity and primer-primer interactions, which are characteristic of other multiplex PCR-based target enrichment methods. However, microdroplet PCR required highly sophisticated, specific instrumentation to run. *See* EX1001.0016 (1:58-63). Further, microdroplet PCR required large numbers of primer pairs when enriching for multiple regions of interest. *Id.*

98. Similar methods were proposed where “several dozen primer pairs can be used together effectively in a multiplexing reaction” by running the reactions in parallel or through microfluidics PCR chips that would allow several thousand singleplex PCRs in parallel. EX2021.0380.

99. Other researchers sought ways to optimize pre-existing multiplex PCR methods by optimizing primer pairs. For example, Shen reported a primer design program called MPprimer, which was designed to address issues with primer specificity for primer pairs. EX2022.0001. Researchers for QIAGEN even published on a “prerequisite for successful multiplex PCR [being] the design of optimal primer pairs,” offering publications to help consumers design optimal primer pairs. EX2023.0005.

2. Capture-by-Hybridization

100. The second major category of enrichment methods was capture-by-hybridization. *See* EX1001.0016 (1:39-47). Like the PCR-based methods discussed above, capture-by-hybridization sought to generate samples containing primarily the genomic region of interest while limiting the presence of extraneous genetic information. Unlike the PCR-based methods, which seek to amplify the signal of the region of interest by copying the region of interest, capture-by-hybridization methods utilize long oligonucleotide chains, termed “probes,” that are designed to hybridize directly to sequences that define, overlap substantially with, or are in close proximity to the region of interest. *See* EX1001.0016 (1:64-66). The probes could

then be selectively isolated, allowing the genomic regions of interest to be physically separated from the unwanted genetic material. *See* EX1001.0016 (1:64-2:4).

101. As capture-by-hybridization relied on target-specific probes to sequester genomic regions of interest, it was generally less susceptible to off-target amplification noise compared to PCR-based enrichment methods. However, capture-by-hybridization typically required larger amounts of input DNA (EX1001.0016 (2:4–10)), as well as more complex probe designs and a lengthy, multi-step workflow for capturing and selecting the target regions. While these probe-based strategies could theoretically cover large panels of genomic sites, practical implementations were often constrained by higher costs associated with synthesizing individual oligonucleotide probes. *Id.* In addition, capture-by-hybridization could exhibit bias toward repetitive elements in the genome—these repeated sequences can cross-hybridize to the capture probes, leading to over-enrichment of repetitive regions and creating unbalanced coverage in the resulting sequencing data. EX2024.1025.

3. Capture-by-Circularization

102. The third major category of enrichment methods was capture-by-circularization. *See* EX1001.0016 (1:39-47). Capture-by-circularization methods utilized linear nucleotide probes designed such that both ends of the probes hybridized at or near adjacent regions in a genomic region of interest. *See*

EX1001.0016 (2:11-16). When properly hybridized, the region between the hybridized regions bound by both ends of the probe could then be copied into the probe while joining the two ends of the probe to create a circle, a process termed “gap fill polymerization.” *Id.* Probe circularization allowed for the selective enrichment of circularized probes by filtering out linear probes and probes that did not circularize properly. Following selection, circularized probes could be re-linearized and the copied region of interest could be further amplified by PCR. *See* EX1001.0016 (2:16-18).

103. Due to the multiple hybridization approach, capture-by-circularization was highly specific for the target region of interest. However, capture-by-circularization probes exhibited a lack of consistent binding to the desired target region of interest, resulting in a variable amount of sequence coverage. *See* EX1001.0016 (2:18-24). Further, the large size of the probes limited the amount of sequencing real estate for the sequences of interest in a given sequencing read, as regions of the probes themselves often took up a sizable portion of the generated sequencing read.

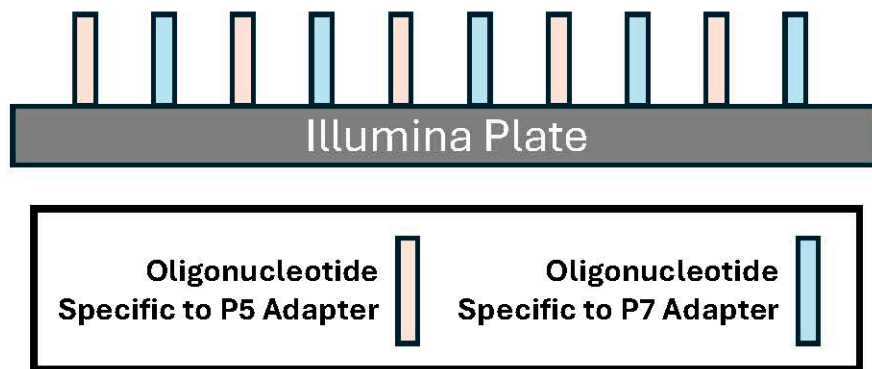
104. As late as 2011, the art recognized that these three target capture approaches suffered from issues of varying degree with “uniformity, efficiency of coverage and off-target capture.” EX2169..0007. Given these issues, practitioners

wondered whether “it may ultimately be more cost-effective to sequence a whole genome” rather than use target capture. *Id.*

I. DNA Hybridization Microarrays and NGS Platforms Rely on Different Technology

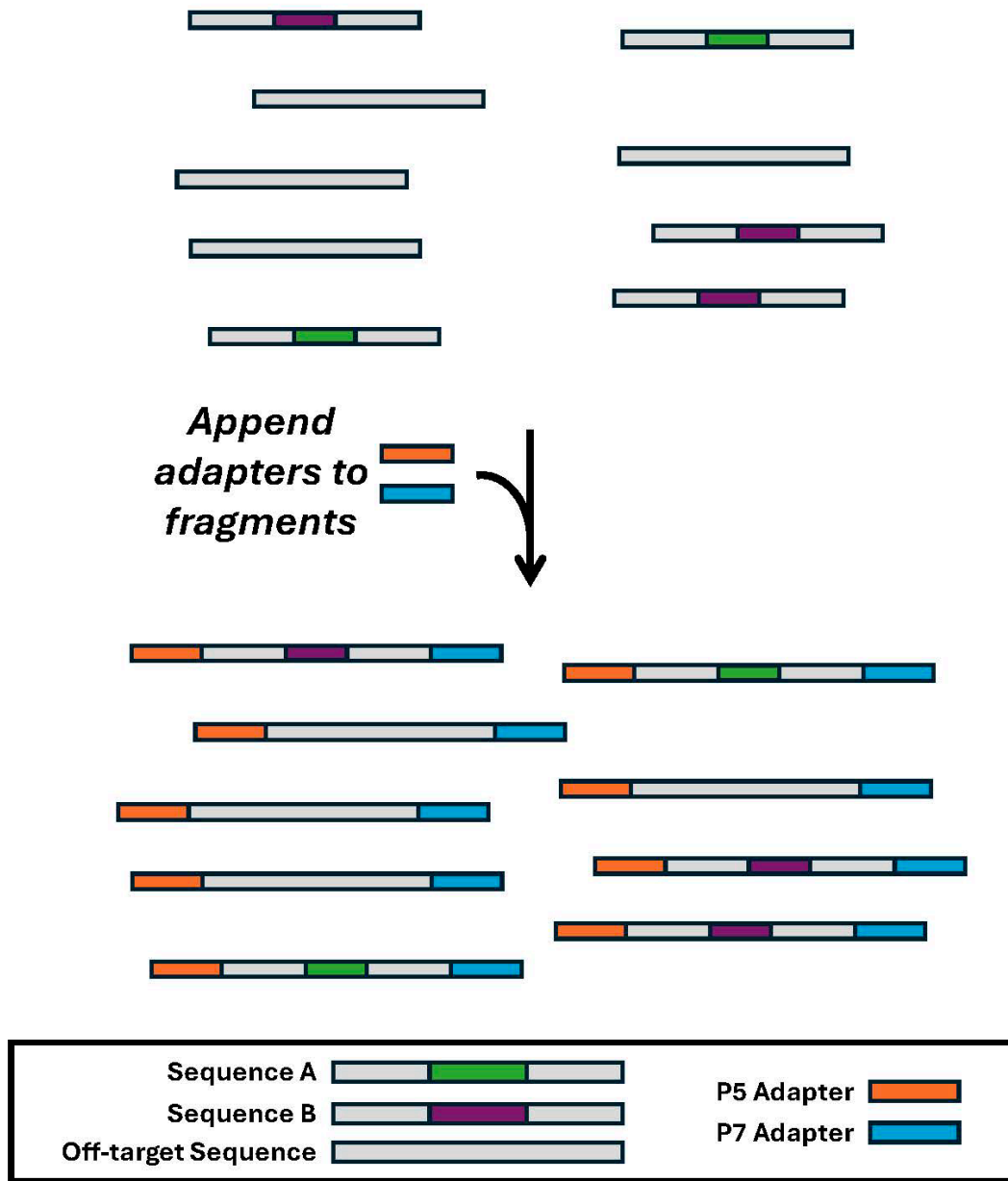
105. As explained above, researchers utilized various methods to understand genetic information, including genotyping through hybridization arrays and NGS platforms. The below examples illustrate the key differences.

106. The surfaces of NGS platforms are designed to bind to universal adapter sequences introduced during sample preparation. EX2007.0336, *see also* EX2030.0002. Taking Illumina as an example, each flow cell contains a dense lawn of two distinct surface-bound oligonucleotides—commonly referred to as P5 and P7 capture oligos—numbering in the millions to billions across the flow cell. These capture oligos are complementary to the adapter sequences (e.g., Illumina’s P5 and P7 primers) ligated to DNA fragments during library preparation, as shown in the image below. By anchoring the prepared DNA fragments to the flow cell surface via these universal adapters, the Illumina platform enables massively parallel sequencing of numerous distinct sequences simultaneously.



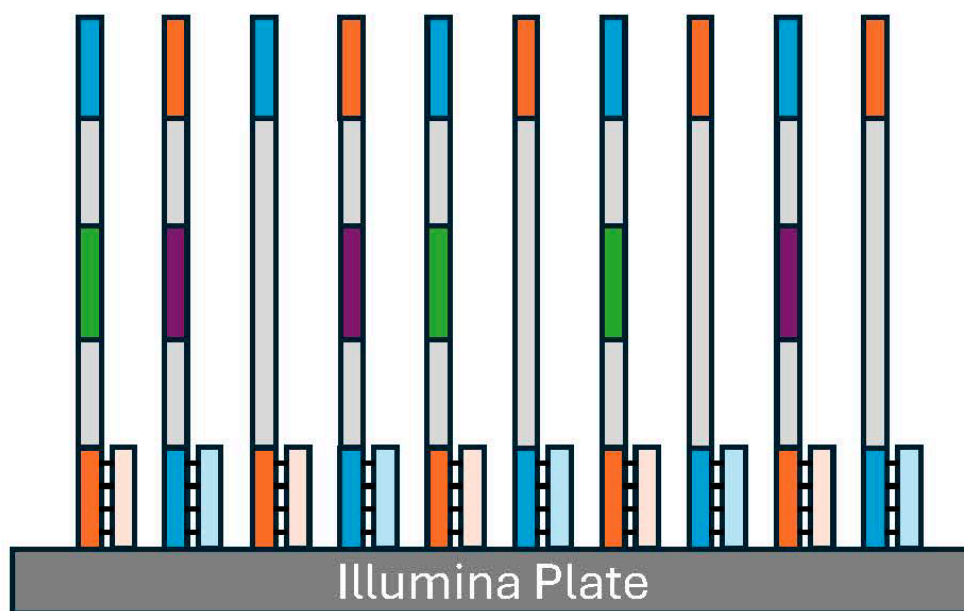
107. If a researcher wanted to sequence their samples with Illumina, the researcher would need to prepare their sample library for sequencing by incorporating the adapters containing the Illumina universal primers at the ends of their nucleic fragments, as shown in the below image.⁶ EX2030.0002.

⁶Note that “Sequence A,” “Sequence B,” and “Off-target Sequence” represent three hypothetical distinct nucleic acid fragments.



108. These samples are then loaded for sequencing, and the adapters containing the universal primers could be used to bind the fragments to the surface

of the sequencer, as depicted below.⁷ EX2007.0004. These same primers could then be used for “bridge amplification”—a process by which sequences hybridized to the surface of the sequencer are amplified on the surface to generate clusters of the same sequence prior to sequencing. *Id.* The P5 and/or P7 primers could then be used as starting points for the Illumina platform to read the fragment’s sequence. EX2007.0003-05, *see also* EX2030.0002.



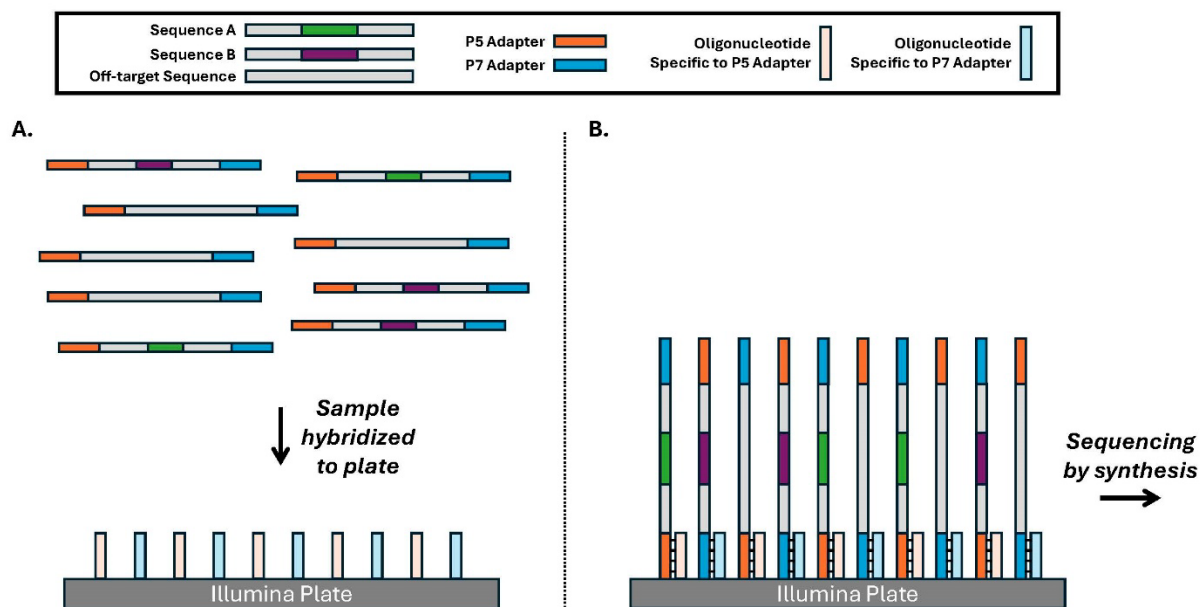
109. Because NGS platforms engaged in massively parallel sequencing, meaning that many different sequences could be sequenced simultaneously, it would

⁷ This graphical representation is a simplified depiction of how sequences are arranged on an Illumina sequencing surface. It does not reflect the full incorporation of sequences by bridge amplification.

have been infeasible to bind samples to the sequencer platform through target-specific probes. In fact, a key feature of the NGS platforms is that the use of common primers enables the binding of many distinct sequences, allowing for massively parallel sequencing. EX2008.0213-14.

110. However, because the surface probes of NGS sequencers are not target-specific, all fragments that contain the universal primer sequences will be sequenced. This generates a significant amount of data about all of the sequences included in the sample. If a researcher was interested in a specific nucleic acid sequence, the researcher would need to prepare sequencing libraries that enrich for the target sequences to avoid the considerable noise from data generated from all the sequences that are not of interest.

Sequencing by Illumina⁸

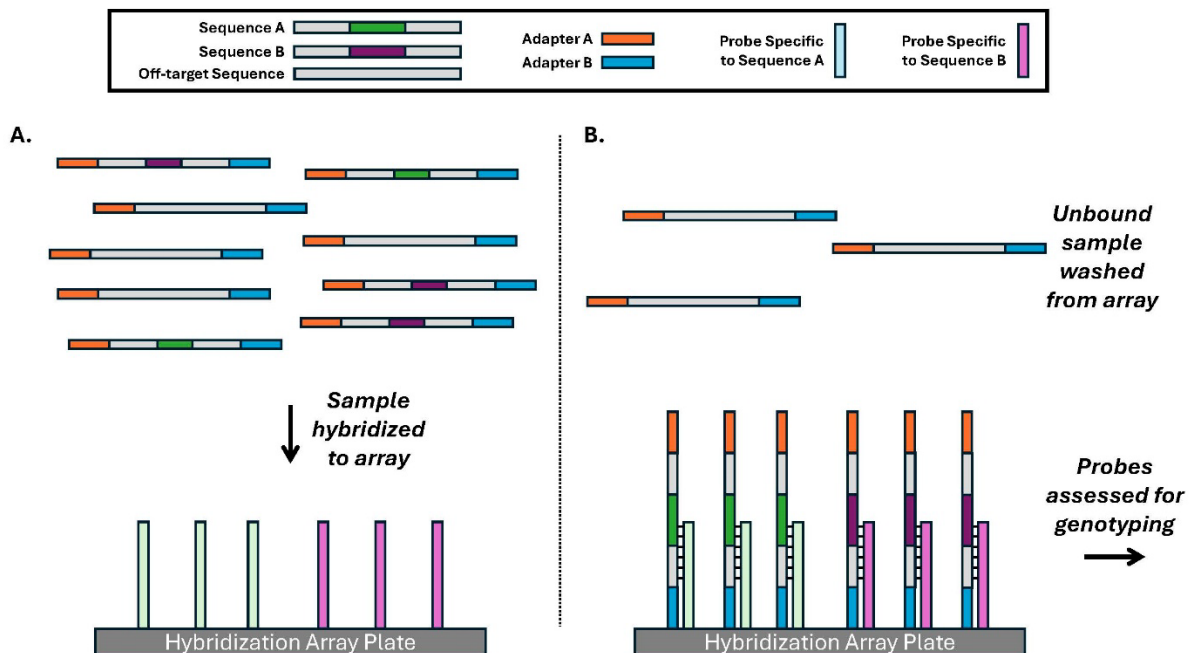


111. In a hybridization array, on the other hand, each probe is highly specific to a target, as discussed above. *See supra*, ¶¶ 65-67. If a researcher was screening a nucleic acid sample for the presence of certain SNPs, that researcher would utilize a hybridization array with probes complementary to the genomic sequence surrounding that SNP (what would be referred to as the genomic locus). EX2004.0011-13, *see also* EX2031.0415, EX2035.1235-36. After the sample was

⁸ This graphical representation is a simplified depiction of the Illumina sequencing process. It does not reflect the full incorporation of sequences by bridge amplification.

prepared, the researcher would run that sample over the array. EX2004.0006. If the locus was present, the nucleic fragment would bind to the probe and produce a detectable signal. *Id.* Through the use of software specific to that array, the researcher could then ascertain whether the locus and SNP were present in the sample. EX2004.0018-19.

Genotyping by Hybridization Array⁹

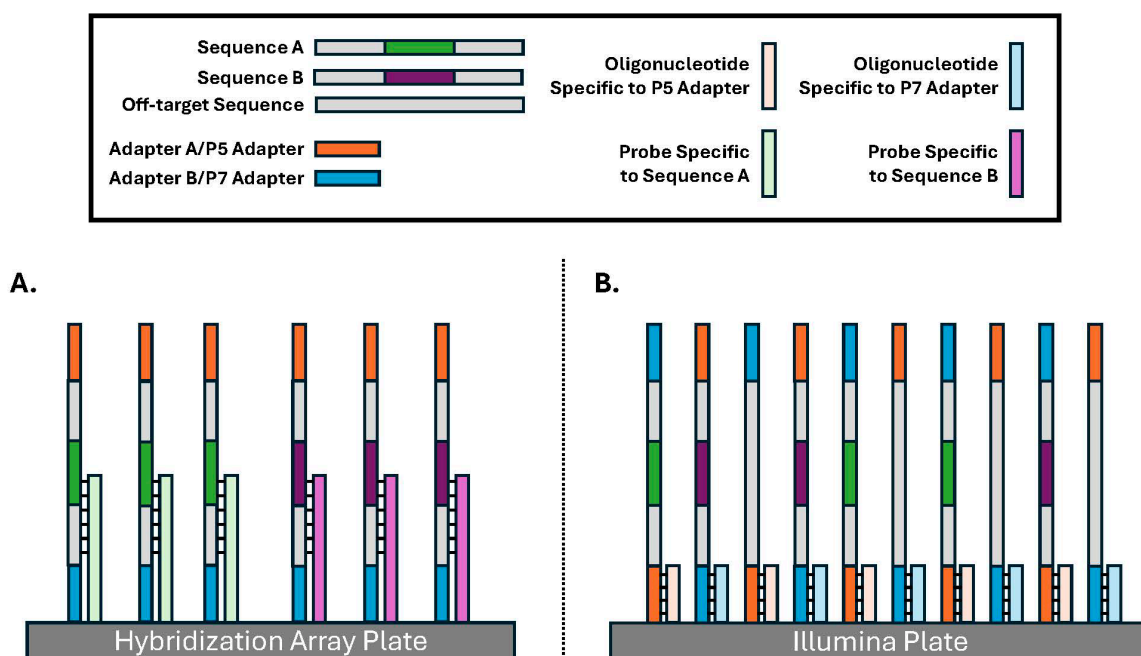


⁹ This graphical representation is a simplified depiction of genotyping by hybridization array. It simplifies the number of distinct hybridization probes present on the array plate. As explained above, a given array plate would usually contain hundreds of billions of probes (including the duplicates of the same probe) as

112. As depicted below, hybridization arrays rely on highly specific probes designed to recognize particular nucleic acid sequences, whereas NGS platforms (e.g., Illumina) depend on universal adapters for binding many distinct fragments.¹⁰

depicted in the graphical representation. Each group of probes would be specific to a particular SNP.

¹⁰ This graphical representation is a simplified depiction of genotyping by hybridization array that simplifies the number of distinct hybridization probes present on the array plate. Further, the graphical representation depicts only the array probes with the target sequence bound at the “0” position. *See supra*, ¶¶ 71-72.



Because hybridization arrays incorporate hundreds of billions of target-specific probes, concerns about off-target hybridization are inherently minimized. In contrast, NGS platforms lack such target-specific surface probes and will sequence any fragment in the library that carries the correct adapter sequences—whether or not the researcher is interested in that fragment. This can result in unwanted or non-relevant reads, increasing both data volume and the complexity of downstream analysis.

V. THE PERSON OF ORDINARY SKILL IN THE ART

113. The Board adopted the following definition in its Institution Decision:

[A] POSA would have had a Ph.D. in molecular biology, molecular genetics, bioinformatics, biochemistry, or a related field. Alternatively, a POSA would have had a B.S. or M.S. degree with three to five years of experience working in any of those fields. Additionally, the POSA would have had experience with NGS sequencing methods and a working knowledge of how those methods functioned.

Paper 7, 18-19 (same for the '108 patent).

114. I agree with the Board that a POSA would possess either a Ph.D. in molecular biology, molecular genetics, bioinformatics, biochemistry, or a related field; or a B.S. or M.S. degree with three-to-five years of experience working in any of these fields. However, it is my opinion that a POSA would have only needed to have routine experience with DNA sequencing methods (including next-generation sequencing which was the dominant mode at the time) and a working knowledge of how these sequencing methods worked in order to troubleshoot experiments—not experience and a working knowledge specific to NGS sequencing.

115. For the purposes of this Declaration, I apply the Board's definition of a POSA. Slight variations on this definition of a POSA would not change my opinions set forth in this declaration, including the definitions adopted by the Board in the related IPR2025-00015 and IPR2025-00016 proceedings.

116. Under either definition of a POSA, I possessed the relevant qualifications as of January 26, 2012. *See supra*, ¶¶ 9-20.

VI. THE PATENTED INVENTIONS

117. Titled “Compositions and Methods for Targeted Nucleic Acid Sequence Enrichment and High Efficiency Library Generation,” the ’012 and ’108 patents are broadly directed to “methods, compositions and kits for targeted nucleic acid sequence enrichment in a nucleic acid sample and for high efficiency nucleic acid library generation for next generation sequencing (NGS).” EX1001.0001; EX1009.0001. The patents are specifically directed to methods “for the selective enrichment of a plurality of defined target sequences from complex DNA with a set of common primers and adaptors, thus circumventing the need for multiplex PCR and multiple primer pairs.” EX1001.0018 (6:12-16); EX1009.0018 (6:14-18). The methods improve on the prior art by creating “a simple, low cost, high throughput

system for target enrichment and library preparation.” EX1001.0018 (6:25-27); EX1009.0018 (6:27-29).

A. The Claims of the '012 Patent

118. The claimed inventions of the '012 patent are defined in 22 claims. The sole independent claim of the '012 patent, claim 1, is reproduced below:

Claim Element	Claim Language
1[pre]	A method for sequencing an enriched nucleic acid sequence of interest, the method comprising:
1[a]	obtaining a nucleic acid fragment ligated to a partial duplex adaptor, wherein the nucleic acid fragment comprises a nucleic acid sequence of interest, wherein the partial duplex adaptor comprises a first adaptor sequence, and wherein the partial duplex adaptor comprises a first strand and a second strand, wherein the first strand is longer than the second strand;
1[b]	annealing one or more oligonucleotides in solution to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the partial duplex adaptor, wherein the one or more oligonucleotides comprise a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;
1[c]	extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the

Claim Element	Claim Language
	partial duplex adaptor with a polymerase, thereby generating one or more oligonucleotide extension products comprising sequence complementary to the first adaptor sequence at a first end, sequence complementary to the nucleic acid sequence of interest, and the second adaptor sequence at a second end;
1[d]	amplifying the one or more oligonucleotide extension products using a first primer that anneals to a complement of the first adaptor sequence and a second primer that anneals at its 3' end to a complement of the second adaptor sequence to enrich for the nucleic acid sequence of interest, wherein products of the amplifying comprise a 3' end with sequence complementary to a sequence on a surface;
1[e]	annealing a strand of the products of the amplifying to the sequence on the surface using the 3' end with sequence complementary to the sequence on the surface; and
1[f]	sequencing the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.

119. The '012 patent contains 21 dependent claims, which are reproduced below:

Claim	Claim Language
2	The method of claim 1, wherein the nucleic acid fragment comprises genomic DNA.

Claim	Claim Language
3	The method of claim 1, wherein the nucleic acid fragment comprises cDNA.
4	The method of claim 1, wherein the nucleic acid fragment comprises RNA.
5	The method of claim 1, wherein the first adaptor sequence and the second adaptor sequence are distinct from each other.
6	The method of claim 1, wherein the one or more oligonucleotides comprise more than one oligonucleotide, and wherein the second adaptor sequence is common to the more than one oligonucleotides.
7	The method of claim 1, wherein the first adaptor sequence, the second adaptor sequence, or both comprise barcode sequence.
8	The method of claim 7, wherein the barcode sequence is at least 8 nucleotides in length.
9	The method of claim 7, wherein the barcode sequence is 12 nucleotides in length.
10	The method of claim 1, wherein the obtaining comprises obtaining a plurality of nucleic acid fragments ligated to a partial duplex adaptor, wherein each partial duplex adaptor of each nucleic acid fragment of the plurality of nucleic acid fragments comprises a barcode, and wherein each barcode is unique.
11	The method of claim 1, wherein the nucleic acid fragment ligated to the partial duplex adaptor is double-stranded prior to step b).
12	The method of claim 11, further comprising denaturing the nucleic acid fragment ligated to the partial duplex adaptor prior to step b), thereby generating a single-stranded nucleic acid fragment comprising the nucleic acid sequence of interest ligated to the first adaptor sequence.

Claim	Claim Language
13	The method of claim 1, wherein the polymerase is a DNA polymerase.
14	The method of claim 1, wherein the first strand of the partial duplex adaptor is ligated to a 5' end of the nucleic acid fragment.
15	The method of claim 1, wherein the second strand of the partial duplex adaptor is ligated to a 3' end of the nucleic acid fragment.
16	The method of claim 1, further comprising fragmenting a nucleic acid, thereby generating a nucleic acid fragment; and purifying the nucleic acid fragment prior to obtaining the nucleic acid fragment ligated to the partial duplex adaptor in step a).
17	The method of claim 1, further comprising performing bridge amplification of the annealed strand after step e).
18	The method of claim 1, wherein the sequencing comprises use of four labeled reversible terminators.
19	The method of claim 1, wherein the sequencing comprises semiconductor sequencing.
20	The method of claim 1, further comprising fragmenting a nucleic acid, thereby generating a nucleic acid fragment; and performing end repair on the nucleic acid fragment prior to obtaining the nucleic acid fragment ligated to the partial duplex adaptor in step a).
21	The method of claim 1, wherein the nucleic acid fragment of the nucleic acid fragment ligated to the partial duplex adaptor in step a) is obtained by performing reverse transcription on an RNA fragment prior to obtaining the nucleic acid fragment ligated to the partial duplex adaptor in step a).
22	The method of claim 1, wherein each end of the nucleic acid fragment is ligated to a partial duplex adaptor.

B. The Claims of the '108 Patent

120. The claimed inventions of the '108 patent are defined in 19 claims. The sole independent claim of the '108 patent, claim 1, is reproduced below:

Claim Element	Claim Language
1[pre]	A method for sequencing an enriched nucleic acid sequence of interest, the method comprising:
1[a]	annealing one or more oligonucleotides in solution in a reaction mixture to the nucleic acid sequence of interest in a nucleic acid fragment, wherein the reaction mixture comprises a plurality of nucleic acid fragments, wherein the nucleic acid fragment comprising the nucleic acid sequence of interest comprises a first adaptor sequence, wherein the one or more oligonucleotides comprise a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;
1[b]	extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment comprising the first adaptor sequence with a polymerase, in the reaction mixture, thereby generating one or more oligonucleotide extension products comprising sequence complementary to the first adaptor sequence at a first end, sequence complementary to the nucleic acid sequence of interest, and the second adaptor sequence at a second end;

Claim Element	Claim Language
1[c]	amplifying the one or more oligonucleotide extension products, in the reaction mixture, using a first primer that anneals to the complement of the first adaptor sequence and a second primer that anneals at its 3' end to a complement of the second adaptor sequence, thereby enriching the nucleic acid sequence of interest, by generating amplified products comprising the enriched nucleic acid sequence of interest; and
1[d]	sequencing the amplified products comprising the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.

121. The '108 patent contains 18 dependent claims, which are reproduced below:

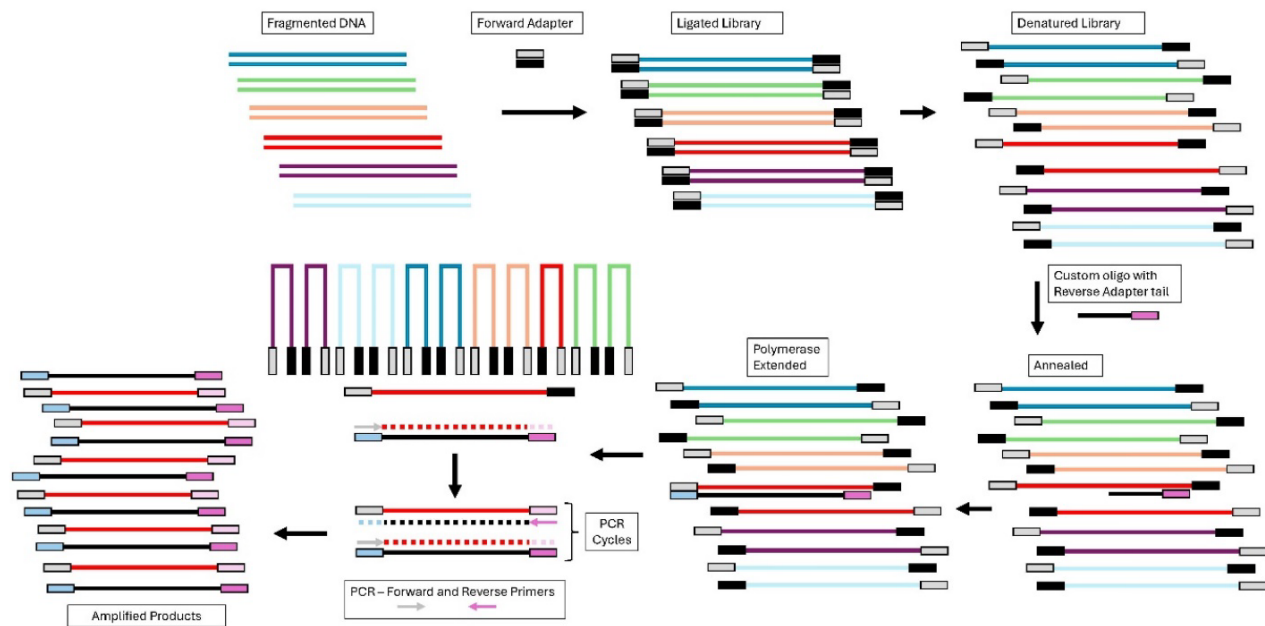
Claim	Claim Language
2	The method of claim 1, wherein the nucleic acid fragment comprises genomic DNA.
3	The method of claim 1, wherein the nucleic acid fragment comprises cDNA.
4	The method of claim 1, wherein the nucleic acid sequence of interest comprises RNA.
5	The method of claim 1, wherein the first adaptor sequence and the second adaptor sequence are distinct from each other.
6	The method of claim 1, wherein the plurality of nucleic acid fragments comprises the first adaptor sequence, and wherein the first

Claim	Claim Language
	adaptor sequence is common to the plurality of nucleic acid fragments.
7	The method of claim 1, wherein the one or more oligonucleotides comprises more than one oligonucleotide, wherein the second adaptor sequence is common to the more than one oligonucleotide.
8	The method of claim 7, wherein the first adaptor sequence and/or the second adaptor sequence comprise a barcode sequence.
9	The method of claim 7, further comprising denaturing double-stranded nucleic acid fragments prior to step a).
10	The method of claim 1, wherein the polymerase is a DNA polymerase.
11	The method of claim 1, wherein the sequencing comprises the use of four labeled reversible terminators.
12	The method of claim 11, wherein the sequencing comprises semiconductor sequencing.
13	The method of claim 1, wherein the sequencing comprises nanopore sequencing.
14	The method of claim 1, wherein each of the plurality of nucleic acid fragments comprises the first adaptor sequence at a 5' end.
15	The method of claim 1, further comprising fragmenting nucleic acids prior to step a), thereby generating the plurality of nucleic acid fragments.
16	The method of claim 1, further comprising ligating the first adaptor sequence to a nucleic acid fragment in the plurality of nucleic acid fragments prior to step a), thereby generating the nucleic acid fragment comprising the first adaptor sequence.
17	The method of claim 1, wherein the amplified products comprise a 3' end with sequence complementary to a sequence on a surface.

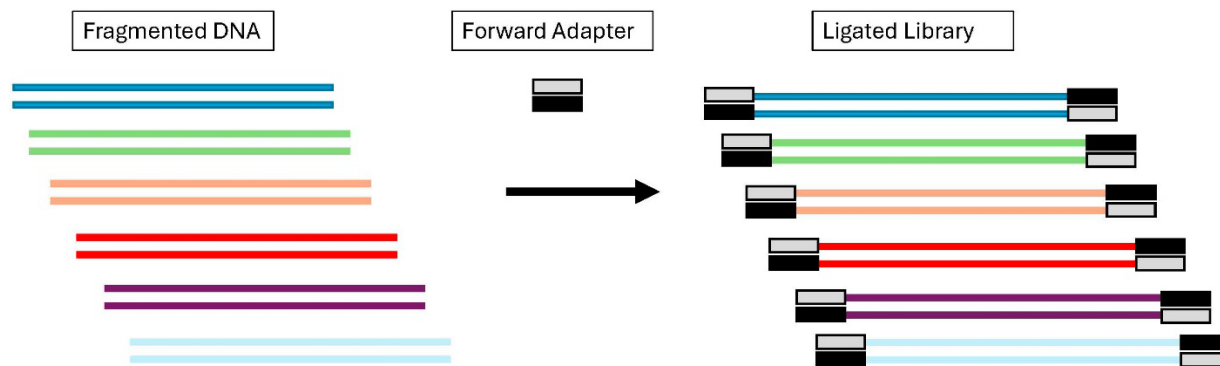
Claim	Claim Language
18	The method of claim 1, further comprising annealing a strand of the products of the amplifying to the sequence on the surface using the 3' end with sequence complementary to the sequence on the surface prior to step d).
19	The method of claim 1, wherein the surface is a surface of a flow cell.

C. Illustration of the Claimed Inventions

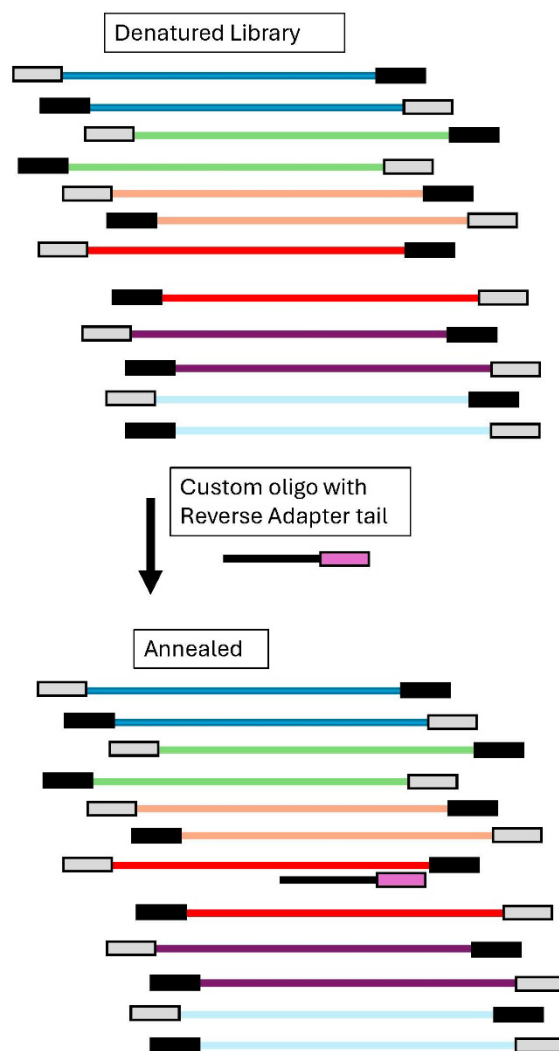
122. I have prepared a series of figures to illustrate how a POSA would understand the claimed limitations. Starting with claim 1 of the '012 patent, a POSA would understand the method is directed to the sequencing of an enriched nucleic acid sequence of interest. The POSA would also understand that many of the claimed steps are directed to library preparation methods for amplicons that are subsequently sequenced with NGS. A representation of the method disclosed in claim 1 is depicted below.



123. In the first step, a POSA would understand that a nucleic acid fragment must be obtained. This nucleic acid fragment must contain a nucleic acid sequence of interest (depicted below as the red fragment). The fragment must also be ligated to at least one partial duplex adaptor (depicted below as the gray and black “forward adapter”), which contains a first adaptor sequence. Ligation can occur at one or both ends of the fragment. While not depicted in the illustration, a POSA would understand that one strand of the partial duplex adaptor is longer than the other, and sits as a single-stranded “overhang.” Because the claim is directed to targeted enrichment of the sequence, a POSA would necessarily understand that this fragment is not alone in the solution, but that instead there are thousands or millions of other fragments in the solution too (the non-red fragments in the depiction).

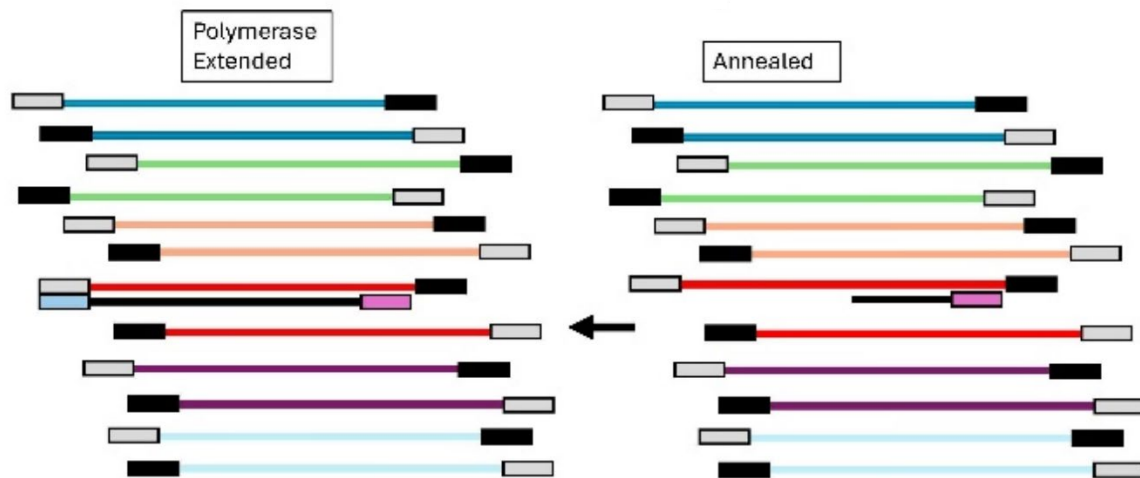


124. In the second step, a POSA would understand that at least one oligonucleotide with sequence complementary to the sequence of interest is annealed to the sequence of interest (depicted below as the “custom oligo with reverse adapter tail”). This oligonucleotide must contain at least 10 bases that are complementary to the sequence of interest (depicted below as the black line). It must also contain a second adaptor sequence that is not complementary to the sequence of interest (depicted below as the pink box).

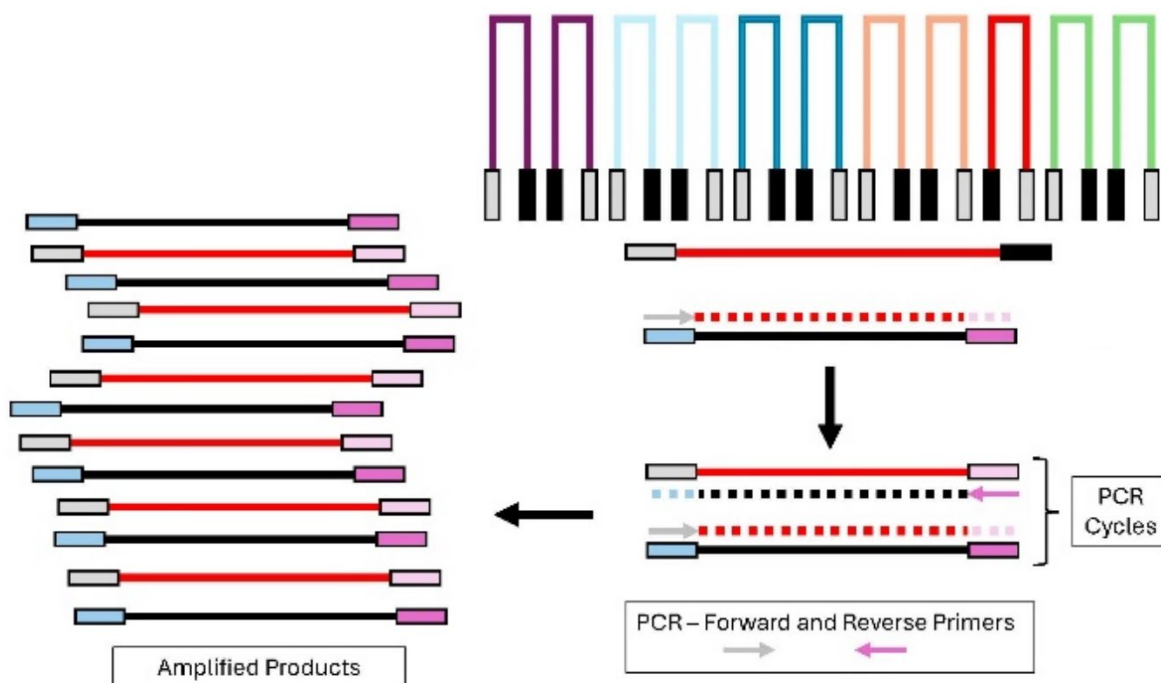


125. In the third step, a POSA would understand that the annealed oligonucleotide is extended, creating an oligonucleotide extension product (depicted below on the left as the longer black strand attached on the left to the blue box and on the right to the pink box). The extension product must contain at least three components: (1) sequence complementary to the first adaptor sequence at one end (depicted below as the blue box); (2) sequence complementary to the nucleic acid sequence of interest (depicted as the black line); and (3) the second adaptor sequence

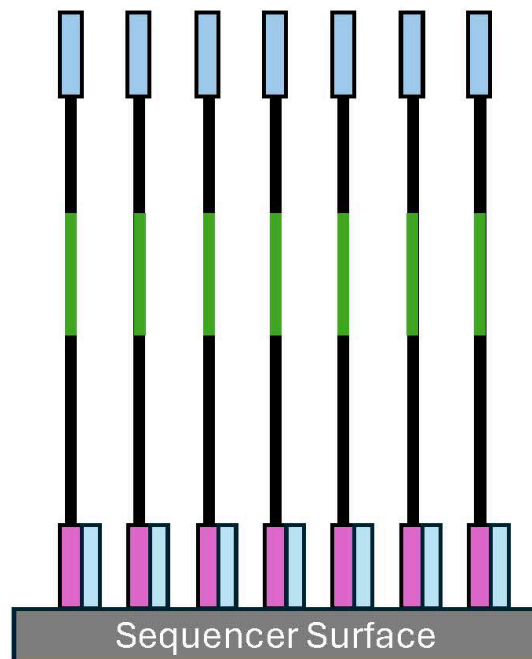
at the other end (depicted below as the pink box). It would be understood that no other fragments but the extension products contain the first and third components.



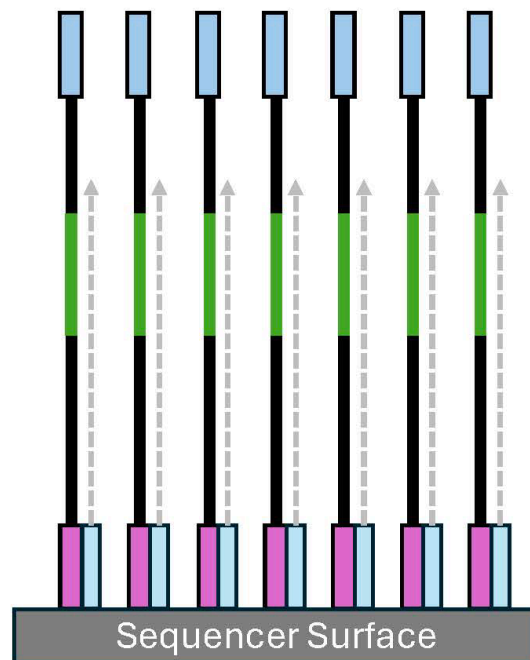
126. In the fourth step, a POSA would understand that the extension products are amplified using two primers: one that anneals to the complement of the first adaptor sequence and another that anneals as its 3' end to a complement of the second adaptor sequence (depicted below as the grey and pink arrows, respectively). Because only extension products with the nucleic acid sequence of interest will have the correct primer binding sites, only the extension products are amplified. This amplification process will necessarily enrich for the sequence of interest by creating a library of amplicons with the sequence of interest and a 3' end with sequence complementary to a sequence on a surface (depicted on the left). Although the figure shows the non-amplified fragments being bent, a POSA would understand that this is meant to show that the fragments are not undergoing amplification and not any specific mechanism, such as PCR suppression.



127. In the fifth step, a POSA would understand that the amplicons are annealed to a surface through their 3' ends. A POSA would also understand that this surface is necessary for fixing the amplicons for sequencing, such as Illumina flow cells.



128. In the sixth and final step, a POSA would understand that the annealed amplicons with the sequence of interest are sequenced with a massively parallel sequencing platform (i.e., NGS). The target sequence is shown in green here to reflect that NGS platforms hybridize to universal adapters, not specific sequences.



129. Claim 1 of the '108 patent discloses a similar method. It is also a method for sequencing an enriched nucleic acid sequence of interest. The first step of this method is to anneal the oligonucleotide to a sequence of interest. As Dr. Metzker acknowledges, this step is identical to the corresponding annealing step of claim 1 of the '012 patent. Compare EX1003, IPR2025-00029, ¶80 to EX1003, IPR2025-00028, ¶80. While the '108 patent specifies that the annealing occurs “in solution in a reaction mixture” wherein the “reaction mixture comprises a plurality of nucleic acid fragments,” a POSA would have understood that the same limitation would necessarily exist for the '012 patent. If the fragment containing the sequence of interest were isolated from fragments not containing the sequence of interest, then it would be impossible to enrich for fragments containing the sequence of interest

(i.e., a solution starting with 100% of sequence X cannot be increased beyond 100%).

130. In the second step, the oligonucleotide is extended. Dr. Metzker acknowledges that this step is identical to the extension step of claim 1 of the '012 patent. Compare EX1003, IPR2025-00029, ¶81 to EX1003, IPR2025-00028, ¶81. As with the '012 patent, this step creates oligonucleotide extension products with (1) sequence complementary to the first adaptor sequence at one end; (2) sequence complementary to the nucleic acid sequence of interest; and (3) the second adaptor sequence at the other end. Because the patent uses the two adaptor sequences to amplify and thus enrich for the sequence of interest, no other fragments in the solution can contain both adaptor sequences at their ends.

131. In the third step, the oligonucleotide extension products are amplified to enrich for the sequence of interest. Dr. Metzker acknowledges that this step is identical to the amplification step of claim 1 of the '012 patent. Compare EX1003, IPR2025-00029, ¶81 to EX1003, IPR2025-00028, ¶81. While this step does not expressly require that the amplicons contain sequence complementary to the second adaptor sequence, a POSA would have understood that the amplification of the oligonucleotide extension products would necessarily result in that outcome. Dependent claim 17 specifies that this sequence must be complementary to a sequence on a surface (with dependent claim 18 further specifying that the amplicons

are annealed to the sequence on the surface prior to sequencing). Additionally, a POSA would have understood that the enriched nucleic acid sequence of interest is based on the population of fragments that contain the two adaptor sequences, which are used to facilitate sequencing. This is because the purpose of the method is to sequence an enriched nucleic acid sequence of interest. Enriching for the sequence of interest but failing to incorporate adaptor sequences that can facilitate sequencing defeats the purpose of the method. Additionally, the patents disclose that the incorporation of these adaptor sequences should occur prior to the enrichment step so as to facilitate an efficient library preparation method. EX1001.0033, cl. 1; EX1009.0033, cl. 1.

132. In the fourth and final step, the amplicons are sequenced with an NGS platform. Dr. Metzker acknowledges that this step is identical to the sequencing step of claim 1 of the '012 patent. Compare EX1003, IPR2025-00029, ¶82 to EX1003, IPR2025-00028, ¶82.

D. Relevant Prosecution History

133. I understand that the patents at issue, the '012 Patent and the '108 Patent, are part of a larger patent family containing additional related patents. I understand that the prior assignee of the patents at issue, NuGEN, filed the first application in the patent family, U.S. Provisional Patent Application No. 61/591,241 on January 26, 2012. EX1001.0001; EX1009.0001. I understand that NuGEN

subsequently filed U.S. Patent Application No. 13/750,768 (“the ’768 application”), on January 25, 2013. EX1001.0001; EX1009.0001. I understand that U.S. Patent Application No. 13/750,768 issued as U.S. Patent No. 9,650,628 (“the ’628 patent”) on May 16, 2017. *See* EX1001.0001; EX1009.0001.

134. I understand that NuGEN filed U.S. Patent Application No. 15/471,785 (“the ’785 application”) as a continuation of the ’768 application on March 28, 2017. EX1001.0001. I understand that the ’785 application issued as the ’012 Patent on July 31, 2018. *Id.*

135. I understand that NuGEN filed U.S. Patent Application No. 16/017,340 (“the ’340 application”) as a continuation of the ’785 application on June 25, 2018. EX1009.0001. I understand that the ’340 application issued as the ’108 Patent on December 29, 2020. *Id.*

136. I understand that the ’012, ’108, and ’628 Patents all claim priority to Provisional Application No. 61/591,241, which was filed on January 26, 2012. EX1001.0001; EX1009.0001.¹¹ I further understand that, for purposes of this

¹¹ I understand that the ’012 and ’108 patents share a common specification. Because of this, I sometimes cite to one of the two specifications herein. To the extent relevant, those references apply equally to the other patent.

proceeding, QIAGEN does not dispute that the priority date as reported on the faces of the '012 and '108 Patents is January 26, 2012. Paper 2, 7.

137. I understand that during prosecution of the applications that would issue as the '012, '108, and '628 patents, the same examiner rejected the proposed claims arguing that they were obvious over two prior art references: U.S. Patent Publication No. 2005/0142577 (“Jones”) and Tucker et al. “Massively parallel sequencing: the next big thing in genetic medicine” (“Tucker”).

138. I understand that the Applicant traversed the Examiner’s rejections, arguing that a POSA would not have had a motivation to combine Jones and Tucker to practice the claimed inventions with a reasonable expectation of success. I understand that as part of the Applicant’s responses, the Applicant submitted a sworn affidavit by Dr. Andrew I. Brooks, in which Dr. Brooks presented evidence and testimony concerning why a POSA would not have been motivated to adapt a method using a hybridization array for NGS and why a POSA would not have had a reasonable expectation of success. I also understand that Dr. Brooks provided evidence and testimony about the same through his accounts of skepticism and unexpected results.

139. Dr. Brooks was widely known and respected in the industry as being at the forefront of genomic sequencing technology. During his decades of work in the field, Dr. Brooks was the Chief Operating Officer of Rutgers University Cell and

DNA Repository (“RUCDR”) Infinite Biologics (in addition to being a long distinguished research professor at Rutgers). At the time, RUCDR was known as the world’s largest university-based biorepositories and one of the largest sequencing laboratories in the country. Dr. Brooks was considered an authority in the field for high throughput sample analysis, having authored numerous publications on the topic. His work included various applied applications to genomic analysis, including diagnostic testing. For example, towards the end of his career at RUCDR, Dr. Brooks develop the first FDA-approved rapid saliva test for COVID-19. The New York Times described this invention as “one of the most reliable means of determining whether someone has the coronavirus.” EX2032.0001. It even prompted Governor Phil Murphy of New Jersey to call Dr. Brooks “one of the state’s unsung heroes.” *Id.*

140. Under either QIAGEN or Tecan’s proposed definitions of a POSA, Dr. Brooks would have easily met the criteria and would have been considered an authoritative expert on high-throughput sequencing technology.

E. CLAIM CONSTRUCTION

141. Dr. Metzker appeared to cast doubt at his deposition on the specific sequencing instruments covered by the claimed limitations. *See* EX2109.0020 (80:1-18). Both patents specifically require sequencing of the enriched sequence of interest “on a massively parallel sequencing platform.” The shared patent

specification explains that “a massively parallel sequencing platform” means an NGS platform. EX1001.0028 (26:64-67). This definition is consistent with how a POSA would have understood the term. *See, e.g.*, EX2169.0001-02.

142. Dr. Metzker appeared to suggest that this definition may encompass hybridization platforms. *See* EX2109.0020 (80:1-18). The shared patent specification treats hybridization platforms as distinct from NGS platforms. *See* EX1009.0016, EX1009.0016. This distinction is also reflected by the Brooks Declaration, where Dr. Brooks distinguished the claimed sequencing of an enriched nucleic acid sequence of interest from genotyping with a hybridization array, such as the Affymetrix microarrays described in Shapero and Jones. Brooks Declaration, EX2026.0003-04. A POSA would have also understood that “a massively parallel sequencing platform” is not a hybridization platform. As an example, various prior art publications distinguished between NGS and hybridization platforms. *See* EX2166.2028, EX2169.0007. A POSA would have considered NGS and hybridization platforms to be distinct. This is consistent with how QIAGEN also treats the two platforms in its Petition. *See* Paper 2, 20.

VII. RELEVANT PRIOR ART

A. Jones (EX1011)

143. U.S. Patent Publication No. 2005/0142577, Methods for Genotyping Selected Polymorphisms (“Jones”) is a patent publication that discloses methods

relating “to enrichment and amplification of a collection of target sequences from a nucleic acid sample and methods of analyzing amplified product.” EX1011.0001, [0002]. Jones names three inventors: Keith W. Jones, Michael H. Shapero, and Stephan P.A. Foder. It is a continuation-in-part of U.S. Patent Application No. 10/264,945 (“Jones 2003”).

144. According to Jones, the claimed inventions are directed to “[m]ethods for genotyping polymorphisms using a locus specific primer that is complementary to a region near a selected polymorphism.” EX1011 at Abstract. Jones discloses that the “invention relates to enrichment and amplification of a collection of target sequences from a nucleic acid sample and methods of analyzing amplified product.” EX1011.0001, [0002]. Jones discloses that “[i]n some embodiments, the invention relates to the preparation of target for array based analysis of genotype.” *Id.*

145. The claims of Jones are directed at methods requiring the use of allele discrimination through hybridization arrays. EX1011.0015 For example, claim 1 requires “hybridizing the amplified extended capture probes to an array of probes comprising allele specific probes for a plurality of the polymorphic positions.” *Id.* As another example, claim 36 requires “hybridizing the amplified fragments to an array of probes comprising probes that are allele specific for allele of each polymorphism in the plurality of polymorphisms.” EX1011.0016.

146. Jones does not disclose any methods of NGS sequencing nor does it make any references to NGS technologies, which were not commercially available at the time.

147. Jones discloses, as a preferred embodiment, the sample preparation method known as “Multiplexed Anchored Runoff Amplification (MARA).” EX1011.0008, [0076]. The MARA method as described in Jones uses locus-specific primers to amplify sequences containing SNPs of interest. EX1011.0008-10. SNPs of interest could be chosen based on several technical criteria, such as whether “the polymorphism is located near a potential priming site (the priming site is where the capture probe hybridizes) for a selected degenerate capture probe, . . . a selected restriction site is located near the polymorphism, . . . and the polymorphism is between the priming site and the restriction site.” EX1011.0011, [0098].

Jones describes that “[t]he amplified product[s] may be analyzed by hybridization to an array of probes.” EX1011.0008, [0076]. Jones describes hybridization arrays as target sequence specific, wherein the probes “hybridize specifically to one or more SNP containing sequences.” EX1011.0008, [0071]; .0013, [0108].

148. Jones discloses that the preferred embodiment of genotyping samples prepared via the MARA method is with hybridization arrays, most preferably using Affymetrix genotyping arrays. EX1011.0007-08, .0013; [0069], [0076], [0077], [0109]. Affymetrix is the assignee of the patent. EX1011 at Cover. The named

inventors of Jones helped to develop Affymetrix genotyping arrays as employees of the company. *See* EX2034.0629. For example, Keith Jones served as the vice president of assay and application product development at Affymetrix. *Id.*

B. Shapero (EX1006)

149. *MARA: a novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays* (“Shapero”) was authored by seven employees of Affymetrix, including the same Michael H. Shapero and Keith W. Jones who were named as co-inventors of the Jones 2003 and Jones publications. Keith Jones was identified as the senior author. EX1006.0001. The article was published in 2004 in the *Nucleic Acids Research* journal. *Id.*

150. As I explain in detail below, Shapero discloses methods for genotyping SNPs using a sample preparation method known as MARA, the same method described as a preferred embodiment of Jones, in conjunction with hybridization arrays. *Id.* Shapero does not disclose any use of NGS methods, platforms, or technologies.

151. As described by Shapero, methods of genotyping in a multiplexed fashion—simultaneous genotyping using multiple primer sets that target different SNPs across the genome—were either technically demanding, limited in their ability to scale beyond 100-plex reactions, or both. EX1006.0001-02. As described above, SNPs are useful markers for studies of genomic function in both health and disease,

including complex diseases in which networks of SNPs can be implicated. Shapero was particularly interested in engaging with emerging “haplotype maps,” efforts to link particular groups of SNPs to health outcomes, which Shapero predicted would “identify a standard set of common-allele SNPs that are expected to provide the framework for new genome-wide studies designed to identify the underlying genetic basis of complex diseases, pathogen susceptibility and differential drug responses.” EX1006.0001. Shapero estimated that “the number of SNPs that need to be genotyped for whole genome association studies using large population-based samples may range from 100,000 to 1,000,000” and therefore “novel target preparation methods need to be developed that, when combined with methods for allele discrimination, allow rapid and accurate genotypic information at a fraction of the cost of current approaches.”¹² EX1006.0002. To these ends, Shapero sought to

¹² A POSA would not have understood Shapero to be assigning weights to the factors or even suggesting that accuracy should be balanced with time and costs. A POSA would have understood that accuracy is the primary concern. Indeed, researchers in the field observed that while the “ease of any given enrichment method and its potential for near-term implementation in a production setting are also considerations,” “the quality of the NGS data produced with a given enrichment

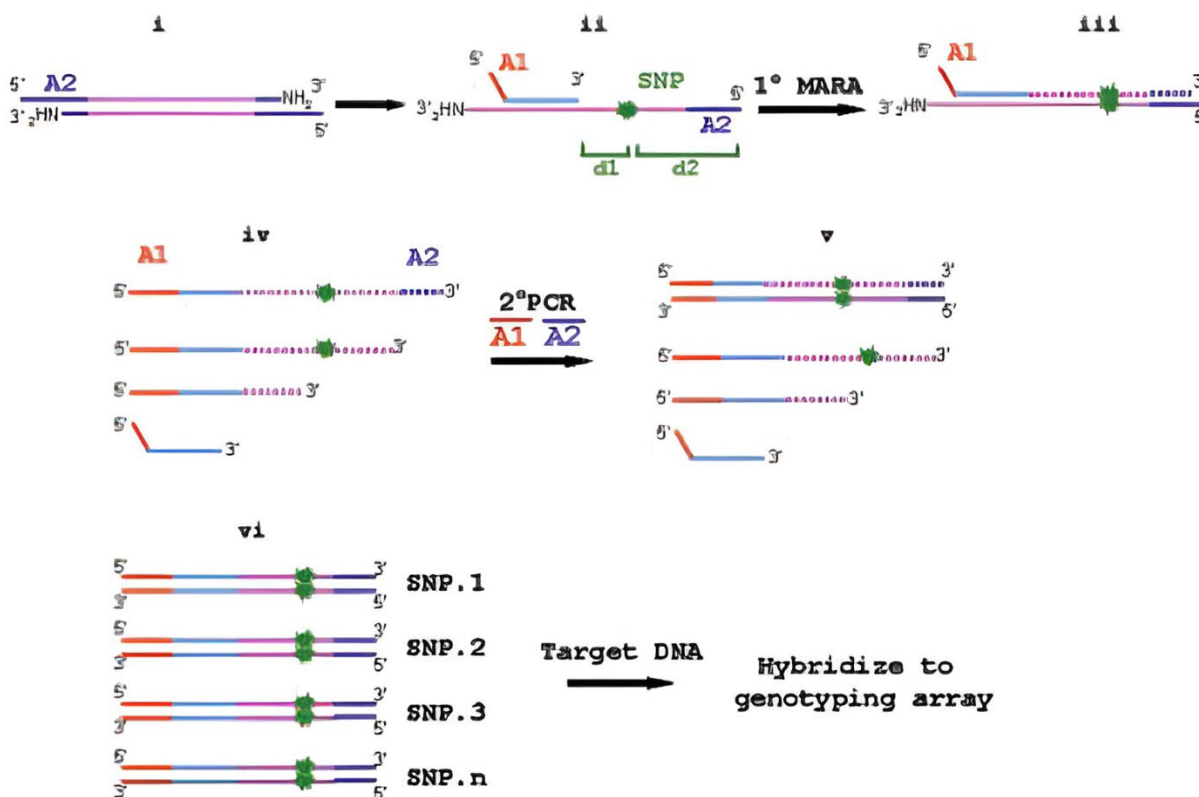
develop “a target preparation method for highly multiplexed locus-specific SNP genotyping that far exceeds the capacity of standard traditional multiplex PCR.” *Id.*

152. Shapero discloses the use of its “novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays” through a series of multiplex reactions. EX1006.0001. The workflow largely remained the same for the various multiplex reactions, except Shapero noted two additions for the 250- and 750-plex reactions. For both the 250- and 750-plex reactions, Shapero compared the results to “[n]egative control reactions” where “no locus-specific primers were used during the primary MARA reaction.” *Id.* For the 250-plex reaction, Shapero also used a digestion enzyme after extension and before amplification. EX1006.0004. Shapero referred to this as “[a]n enrichment step.” *Id.* I discuss this step further in the next section.

method on a given sample type will ultimately dictate the strategies we elect to further optimize.” EX2063.0448.

1. The MARA Method

153. Shapero discloses the “Multiplexed Anchored Runoff Amplification (MARA)” method. MARA is disclosed in Figure 1, reproduced below.

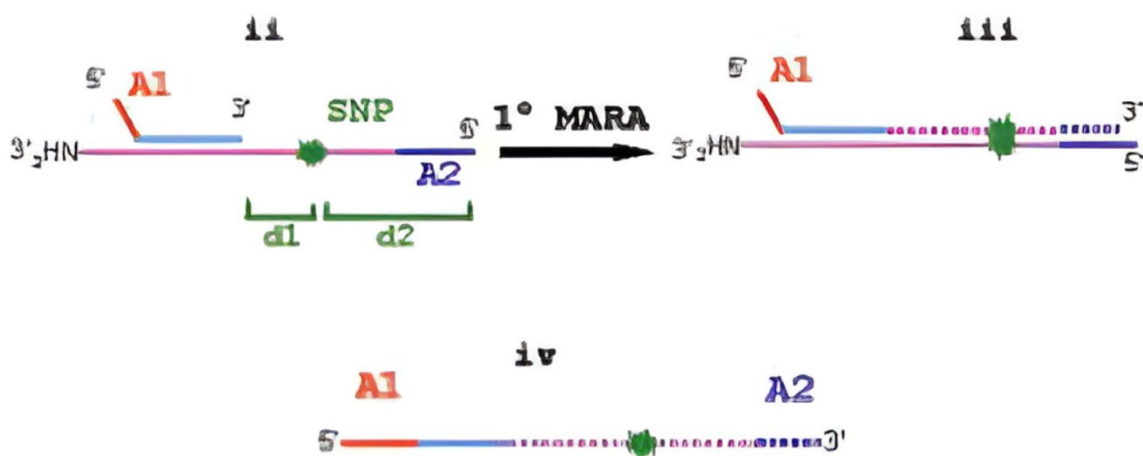


EX1006.0004.

154. MARA relies on target-specific primers designed to bind near the loci of interest containing the SNP site. DNA is first fragmented, and “[e]ach biallelic nucleotide is targeted with a single oligonucleotide primer that contains common sequence at the 5' end (designated A1) followed by locus-specific sequence (~25 bases) at the 3' end.” EX1006.0003. Adapters containing the common sequence

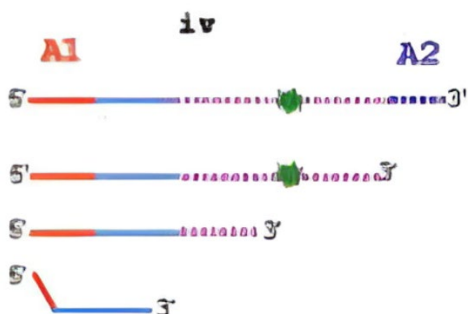
(designated A2) were added in a non-specific manner, and as such all DNA fragments acquire adapters at both ends of the fragment. EX1006.0002-03.

155. The design of the A1 oligonucleotide primers were “constrained by several parameters” including that “(i) probe sequences tiled on the array (16 bases flanking the SNP on either side) [were] excluded from use in the primers and must not contain the restriction enzyme recognition sequence used to digest genomic DNA in MARA” and “(ii) the total length of the extension product . . . [must be] >100 bases and <750 bases.” EX1006.0003. The single oligonucleotide primers “anneal[ed] to the target and [were] elongated during the first cycle” such that “the adapter sequences ligated to the genomic DNA [were] incorporated into the new extension product provided that elongation [was] complete through the end of the restriction fragment.” *Id.* A representative example of the extension step and extension product is shown below.



EX1006.0004.

156. Shapero discloses that this extension process occurs over “multiple [PCR] cycles of primer annealing and extension.” *Id.* Shapero discloses four variations of products that can result as part of the extension process:



Id. Starting from the top and working down, Shapero depicts (1) a full extension product containing both primer binding sequences, (2) a partial extension product containing the A1 primer site and SNP, (3) a partial extension product containing the A1 primer site but no SNP, and (4) the oligonucleotide. For the 250-plex reaction, Shapero discloses that a single round of PCR occurs at the end of the extension process using the T7 primer, which binds to the A2 binding site. EX1006.0003-04. Extension of this primer creates double-stranded DNA for every fragment containing the A2 binding site, i.e., every fragment in solution with the ligated adapters. *Id.* The bottom three products in the above depiction do not contain the A2 binding site. In a subsequent reaction, Shapero introduces Exonuclease I, which is a digestion enzyme that removes any single strand products from the mixture. As Shapero explains, “[t]his step enriches for full-length products generated by the primary reaction.” EX1006.0004.

157. But Shapero goes on to explain the limitations of the digestion process: it “cannot discriminate between products derived from specific and non-specific primer-target DNA annealing.” *Id.* A POSA would understand this to mean that the digestion enzyme could not remove off-target extension products that were fully extended. As all DNA fragments contained adapters, any fragments inadvertently bound by a primer due to off-target primer binding would also result in an extension product.

158. For the 750-, 1100-, and 4365-plex reactions, there was no comparable “enrichment” step with the digestion enzyme, so full off-target products as well as the various forms of incomplete extension products and primers would be carried into the next set of PCR reactions.

159. Following the extension step, “[a] secondary amplification step [was] then carried out using an aliquot of the primary MARA reaction as template.” EX1006.0003. The secondary amplification step utilized “[u]niversal primers” that “anneal to common sites present in all full-length amplification products” such that “secondary amplification is independent of the sequence of the individual locus-specific amplicons generated in the primary MARA reactions.” *Id.* As such, all extension products, regardless of target specificity, were subsequently amplified during the secondary amplification step.

160. The population of amplified DNA fragments was then assessed by genotyping via hybridization array. EX1006.0003-04. As previously described, hybridization arrays contain hundreds of billions of oligonucleotide probes designed to be complementary to the DNA fragments containing the SNPs of interest. As such, only amplified DNA fragments with sequences complementary to the array probes are bound by the array and subsequently genotyped. Amplified DNA fragments not specific to the array probes, including fragments generated by off-target primer binding, are not bound by the array. Instead, those fragments are washed from the array following hybridization and are not reflected in the subsequent genotyping. Following hybridization, the hybridized arrays were stained and scanned to capture the resulting probe signal, and algorithms was employed to make genotyping calls. *Id.*

161. As the MARA workflow was designed for genotyping via a hybridization array, “MARA [did] not require absolute specificity of primer annealing and elongation since target recognition and allele discrimination is achieved through hybridization to high-density arrays.” *Id.* Shapero recognized that “[w]ith MARA, allele discrimination is achieved by hybridization to arrays containing SNP-specific complementary sequences rather than a universal array which requires the use of tags or zip codes attached to the primers for genotype deconvolution.” EX1006.0007. Thus, according to Shapero, “the requirement for

absolute specificity of primer annealing to target sequences in genomic DNA is alleviated.” *Id.* Shapero further predicted, based on “[t]he theoretical sequence complexity of target DNA generated by MARA” and the genotyping capacity of high-density sequencing arrays, that “even if the primers contribute to non-specific amplification products, the complexity should not reach a level that would affect array hybridization.” *Id.* Indeed, Shapero made clear that it was the combination of MARA “with a method for highly parallel allele discrimination, such as high-density DNA oligonucleotide array hybridization, [that] afford[ed] a level of locus-specific SNP genotyping that [was] not readily accessible through traditional target preparation methods such as multiplex PCR.” *Id.* A POSA would have understood Shapero to mean that the microarrays are enabling this “level of locus-specific SNP genotyping” and that because of the microarrays, it is unnecessary to address the target specificity and off-target binding issues caused by PCR amplification.

2. Shapero Validates MARA Using Genotyping Array Results

162. In order to assess the viability of MARA, Shapero used the method to genotype different samples using multiple different arrays. In order to assess the quality of the genotyping generated using MARA, Shapero employed a defined scoring system. EX1006.0004-05. Specifically, Shapero reported the genotyping results as (1) a “discrimination score” for a given probe pair, wherein meeting a threshold score “serves as a measure of detection of DNA target sequences,” (2) the

“call rate” (the “percentage of successful genotype calls within a DNA sample”), and (3) the “concordance” (a “comparison of MARA genotype calls with WGS [(whole genome sampling assay)] genotype,” which in Shapero served as a metric for consistency between MARA genotype calls and previously published WGS genotype calls). *Id.*

163. Shapero first validated MARA using a 250-plex MARA assay. *Id.* Primers were designed to target known SNPs that exhibited “reproducible high call rates” so that “genotype call rates [could be used] as a surrogate for the presence (successful primer) or absence (unsuccessful primer) of MARA-generated target DNA.” EX1006.0004. DNA samples from nine distinct individuals were genotyped using the MARA protocol. *Id.* Because “[a] small number of primers target amplicons containing more than one SNP tiled on the array,” Shapero predicted that 278 SNPs would be present in the amplified products assessed by hybridization array. *Id.* Shapero used a “predecessor array” that “used the same probe sequences and tiling strategy to generate genotype calls as the commercially available Affymetrix GeneChip 10K Mapping Xba_131 Array.” EX1006.0003.

164. Assessing the SNPs via hybridization array, “99.4% of the SNPs targeted by locus-specific primers across nine individuals had a mean DS > 0.08,” indicating the bulk of the array probes “detect[ed their corresponding] DNA target

sequences.” EX1006.0004-05. Shapero further reported a mean call rate of 96.2% and a mean concordance of 98.6% across the nine DNA samples.¹³ EX1006.0005.

165. Shapero compared these results to control samples where the MARA assay was conducted, but the locus-specific primers were omitted, meaning that no extension or amplification of the DNA sample took place. *Id.* When the MARA assay was conducted in the absence of the locus-specific primers, only 6.6% of the SNPs had a mean discrimination score above the threshold. *Id.* Similarly, in the

¹³ Concordance rates are a common way in the industry to report how one instrument performs in genotyping a sample compared to reference genotypes. Affymetrix published that the standard concordance for the GeneChip Mapping 10K Xba 131 Array, which Shapero identified as the analogous array to the predecessor array used in Shapero, is greater than 99.6%. EX2033.0002. A POSA would not read the difference between 98.6% for the 250-plex results in Shapero and 99.6% as a simple difference of 1%. Instead, a POSA would consider these numbers based on the percentage of discordance, which is an indicator of error. Affymetrix reported a 0.4% discordance rate for the general use of the array. Through the method reported in Shapero, Shapero reported a 1.4% discordance, which is almost four times higher than standard discordance. This means that the arrays, as used in Shapero, exhibited higher error rates than standard for the array.

absence of MARA preparation, Shapero reported a mean call rate of 6.3%, and noted that “only five of these SNPs were called across at least five of the nine DNA samples.” *Id.* This comparison demonstrated that “that the presence of locus-specific primers allows specific target sequences to be amplified, and in the absence of primers, only small quantities of non-specific products are generated.” *Id.* In other words, the absence of locus-specific primers impaired the amplification of the entire sample.

166. Shapero then “increased [the multiplexing level] above 250 primer-plex by the addition of 500 primers to the original MARA 250-plex reactions,” raising the total multiplex reaction to 750 primers. *Id.* The array was designed such that only 500 primers generated targets that could hybridize to the array, evaluating 535 SNPs. *Id.* The 750-plex MARA assay resulted in 99.0% of the SNPs achieving a mean discrimination score above the threshold. *Id.* Conversely, the controls indicated that 4.0% of SNPs achieved a mean discrimination score above the threshold when the primers were removed from the MARA assay. *Id.* Similarly, the 750-plex MARA assay achieved a mean call rate of 95.4% and a mean concordance of 98.7% (1.3% discordance) across the nine DNA samples, while the same assay in the absence of the locus-specific primers resulted in a mean call rate of 3.6%. *Id.*

167. Finally, “[i]n order to assess higher levels of multiplexing,” Shapero designed “a new array [] containing 8936 SNPs from human chromosome 21.”

EX1006.0005-06. Shapero focused on a “total of 4365 oligonucleotides [locus-specific primers] . . . targeting a total of 4978 SNPs tiled on the array.” *Id.* Using seven DNA samples, Shapero prepared genotyping libraries using these primers via MARA in two ways as, (1) a single bulk reaction containing the 4365 locus-specific primers and, (2) by randomly pooling the locus-specific primers into four groups consisting of 1113, 1140, 1056 and 1056 locus-specific primers prior to the MARA assay (referred to as the “1100-plex(x4)” group). *Id.* Both preparations were then assessed via the designed hybridization array. *Id.* Shapero applied a filter system to the results to group them into “pre-filtered” and “post-filtered” groups, wherein “post-filter subsets of the 4978 SNPs were identified using selection criteria that required at least a 70% (5/7) call rate across the seven DNA samples.”

EX1006.0006. In order to assess concordance, Shapero compared the genotype results to public data from Perlegen Sciences. *Id.* Unlike the prior 250- and 750-plex assays, Shapero did not report discrimination scores. I have prepared a table summarizing this filtering below (see EX1006.0005-07 for Shapero’s reporting of the data).

Samples	SNPs	Original SNPs (%)	Call Rate (%, Average)	Concordance (%, Average)
Bulk Reaction (Pre-filter)	4976	100	86.4	98.1
Bulk Reaction (Post-filter)	3908	78.5	96.8	99.1
1100-plex(x4) (Pre-filter)	4976	100	84.2	98.5
1100-plex(x4) (Post-filter)	3807	76.5	96.0	99.2

168. As shown in the table above, approximately 25% of the SNPs were not called with enough consistency across samples to meet the 70% call rate threshold. *Id.* A POSA would have understood that this was a typical threshold for which to determine whether the target sequence is present in the solution. When the below-threshold SNPs were included in the analysis, the overall SNP call rate was 86.4% and 84.2% for the bulk and 1100-plex(x4) samples, respectively. *Id.* When the filter parameter was applied and the lower call rate SNPs were excluded, the average call rates increased to 96.8% and 96.0%, respectively. *Id.* Without filtering, the concordance rates were 98.1% (1.9% discordance) and 98.5% (1.5% discordance), respectively. When compared to the array's standard concordance rate of over 99.6% (0.4% discordance), this data showed error rates four-to-five times higher than the standard. From this experiment, Shapero concluded that "at both multiplex levels, over 3500 independent primers are generating specific target sequences" and that "~75% of the input primers function at multiplex levels >1000 bp." EX1006.0006-07.

169. Shapero evaluated several parameters in an attempt to characterize the approximately 25% of calls that fell below the call rate threshold, noting “[t]here were no observable differences between the groups with regard to G:C% of the primers, G:C% of the probes tiled on the array, and G:C% and length of the Sau3AI genomic DNA restriction fragments,” suggesting G:C% was not a factor.¹⁴ EX1006.0006. Shapero disclosed there was “a difference in the length of the extension products, with a higher relative percentage of the SNPs belonging to the discordant group present on extension products in the range of 600–800 bp,” but did not ascribe this difference as leading to a lower call rate. EX1006.0006-07. Shapero further noted that “the overall primer conversion rates do not differ significantly between 1100 primer-plex (x4) and 4365 primer-plex, suggesting that non-specific primer-primer interactions at the higher multiplex level may not solely be responsible for failed target generation.” EX1006.0007. Ultimately, Shapero did not provide an explanation for the failed target generation. At best, the Shapero data confirmed that the MARA method would work with a hybridization array; however, Shapero did not address how this method may be used for other purposes and did not provide explanations for the higher error rates.

¹⁴ A POSA would have understood that differences in G:C% were known to strongly affect hybridization.

C. Syvänen (EX1012)

170. Syvänen is a 2001 review article concerned with different methodologies for SNP genotyping. Syvänen pre-dates the advent of NGS platforms and does not describe any NGS methods or platforms. Syvänen discusses the role of SNP analysis in studies of disease pathology and as a method of clinical diagnostics. Syvänen “review[s] the current state of the art of SNP-genotyping technology, with an emphasis on how amenable the different methods are to multiplexing to increase throughput and bring down the costs of the assays.” EX1012.0931. Syvänen describes SNP-genotyping methods as underpinned by at least one of six biochemical processes. EX1012.0932. Syvänen further characterizes the features of several SNP-genotyping methods as shown below.

Table 2 | **Features of SNP-genotyping methods**

Method	Most significant advantage	Disadvantage	Special feature
Hybridization methods			
Reverse dot blot ²⁸	Multiplexing possible	Limited genotype discrimination	Precursor to microarrays
GeneChip microarrays ³⁹	Very high probe density	Prone to non-specific background	High-tech manufacturing process ⁴⁰
DASH ⁴⁴	Inexpensive labelling method	Complex design rules	Dedicated instrument
PNA ⁴⁶ and LNA ⁴⁸ probes	Specific as ASO probes	Not widely available	Innovative chemical design
TaqMan ^{49,50}	Simplicity of assay	Expensive probes	Quantitative real-time PCR
Molecular Beacons ^{51,52}	Simplicity of assay	Expensive probes	Versatile stem-loop structure
Allele-specific PCR			
Real-time PCR assay		Requires optimization	
Intercalating dye ⁶⁰	Inexpensive labelling method	Non-specific products detected	Accurate quantification ¹³⁸
FRET primers ^{61,62}	Simplicity of assay	Expensive primers	Universal FRET primers
AlphaScreen ⁶⁴	Detects specific product	Expensive probes	Applicable to ASO hybridization
Primer extension			
Accurate genotyping		Similar reagents as in PCR	
SNPstream, GBA ⁶⁸	Inexpensive and robust	Multiple detection steps	Automated high-throughput process
Multiplex minisequencing, SNaPshot ^{69,71}	Multiplexing capacity	Size separation step	Compatible with capillary DNA sequencers
Pyrosequencing ^{72,74}	Sequencing of up to 50 bases	Expensive, difficult to multiplex	Dedicated instrument
MassEXTEND, MassArray ⁷⁵	Labelling method avoided	Expensive instrument	Multiplexing capacity ⁷⁶
GOOD assay ⁷⁷	Labelling method avoided	Multi-step procedure	Sensitive mass spectrometric detection
Microarray miniseq, APEX ^{37,70}	Potential for high throughput	Requires microarray instruments	Four-colour or single-colour detection
Microarray primer extension ⁸²	Potential for high throughput	Requires microarray instruments	'Array of arrays' format
'Tag' arrays ^{85,86}	Flexible assay design	Requires microarray instruments	Generic 'tag' microarrays
Coded microspheres ⁸⁷	Multiplexing potential	Microspheres not widely available	Flow cytometric detection
TDI, fluorescence polarization ⁹⁴	Simplicity of assay	Difficult to multiplex	Universal detection principle
Oligonucleotide ligation			
Variety of assay formats		Multiple labelled probes required	
Colorimetric OLA ^{95,96}	Robust assay	Multiple detection steps	Resembles ELISA
Sequence-coded OLA ⁹⁷	Multiplexing capacity	Size separation step, expensive probes	Polymer for sequence coding
Microarray ligation ¹⁰⁰	Potential for high throughput	Requires microarray instruments	Universal zip code microarray
Ligase chain reaction ¹⁰²	Alternative to PCR	Lower efficiency than PCR	Thermostable ligase
Padlock probes ¹⁰⁸	Localized detection	Probes difficult to produce	Avoids target amplification
Rolling circle amplification ⁹⁸	Signal amplification	Steric hindrance on solid phases	Based on circularized oligonucleotides
Endonuclease cleavage			
Restriction site analysis	Inexpensive, requires no equipment	Not suitable for high throughput	Traditional method
Invader assay ¹¹²	PCR amplification avoided	Requires large amount of DNA	Interesting FLAP endonuclease

APEX, arrayed primer extension; ASO, allele-specific oligonucleotide; DASH, dynamic allele-specific hybridization; ELISA, enzyme-linked immunosorbent assay; FRET, fluorescence resonance energy transfer; GBA, genetic bit analysis; LNA, locked nucleic acid; OLA, oligonucleotide ligation assay; PNA, peptide nucleic acid; TDI, template-directed incorporation.

EX1012.0934.

171. One of the SNP-genotyping methods disclosed by Syvänen is “mini-sequencing,” the term used by Syvänen to describe “methods based on single nucleotide primer extension”—SNP-genotyping methods that use nucleotide incorporation via a polymerase to identify SNPs of interest. EX1012.0935. Syvänen discloses that mini-sequencing methods utilize primers specific to target sequences

containing the SNPs of interest such that extension from these primers incorporates nucleotides that reflect the presence of different nucleotides at the SNP location. EX1012.0935-38. Syvänen discloses various methods by which the incorporated nucleotide, reflective of the SNP of interest, can be detected, including fluorescent labels and measuring molecular mass. *Id.*

172. Syvänen further describes how the aspects of sequencing machines could be applied to “homogeneous allele-specific PCR,” a class of genotyping methods wherein “PCR primers with the 3'-end complementary to either of the nucleotides of a SNP can be used in combination with a common reverse PCR primer to selectively amplify the SNP alleles.” EX1012.0935. Syvänen discloses that capillary channel systems such as “96-channel capillary sequencing instruments, microplate array diagonal electrophoresis or capillary array electrophoresis microplates” could be used to achieve high-throughput size separation of PCR products. *Id.* As an example, Syvänen cites to Applied Biosystems’s SNaPshot, which was an older “mini-sequencing” platform that analyzed SNPs through single nucleotide extensions and electrophoresis.

173. Syvänen notes that “without a consecutive target-specific detection reaction or size-separation step, the specificity of the [homogeneous allele-specific PCR] method might be hampered owing to primer dimers and other spurious amplification products that will not be distinguished from the actual PCR products,”

and that “[a] limitation of all variants of allele-specific PCR is that the reaction conditions or primer design for selective allele amplification must be optimized empirically for each SNP.” EX1012.0935.

D. Delseny (EX1007)

174. Delseny is a 2010 textbook review chapter that discusses NGS platforms as of 2010. EX1007.0407-09. Delseny further describes the application of NGS technology to plant biology including efforts to sequence several different plant species, and the broader application of sequencing methods to address biological questions. EX1007.0408. Delseny does not address sequencing library preparation methods in any detail and provides no mention of single-primer enrichment strategies.

175. Delseny “review[s] some of the next-generation sequencing technologies (NGST),” including the 454 sequencer, Illumina platforms, Ion Torrent, and PacBio platform. EX1007.0408-11. For each platform Delseny highlights, Delseny provides a summary of the sequencing technology utilized by that platform. *Id.* Delseny describes “[t]he major advantage of all the NGSTs is that their throughput is much higher than that of classical sequencing.” EX1007.0411. However, Delseny notes that “all these technologies have a number of limitations that, at least initially, restricted their use to some applications” including short read length, sequencing of repeat regions, bias introduced during sequencing library

construction, and “sequencing errors and artefacts” *Id.* Delseny further notes that each sequencing method “ha[s] its own weakness,” noting as an example in the case of the 454 sequencing that “[m]istakes occur with short homopolymers with the 454 technology because multiple incorporation of the same nucleotide instead of single incorporation can occur at each round.” *Id.* As such, Delseny considered Sanger sequencing to have “the highest accuracy . . . and remain[] the ‘gold standard.’” *Id.*

176. Delseny provides a high level overview of NGS technology, and does not concern itself in any detail with the particulars of sequencing library preparation. At most, Delseny describes sequencing library preparation as “remain[ing] tricky and involv[ing] several steps of fragmentation, adapter ligation and PCR amplification,” wherein “[e]ach step can introduce representation bias and artifacts, which must be carefully controlled for.” *Id.*

177. Delseny provides no mention of any single-primer enrichment technology, nor does it provide any disclosure as to how to address off-target primer binding. *See id.*

1. Delseny Does Not Discuss the Use of a Single Target-Specific Primer for Targeted Sequencing

178. Delseny’s discussion of “targeted sequencing” is similarly limited. Delseny provides a brief, high-level discussion of efforts to sequence particular regions of the genome. EX1007.0417-18. Again, Delseny provides no disclosure of any single-primer enrichment technology or any disclosures concerning how to

address off-target primer binding. Indeed, the only method of target capture expressly disclosed by Delseny is “immunoprecipitating chromatin with specific antibodies and analyzing select regions of the genome] by extracting and sequencing the corresponding DNA,” otherwise known as chromatin immunoprecipitation, a protein capture technique wholly distinct from single-primer enrichment. EX1007.0418.

179. I have reviewed the specific citations referenced by Delseny in the “targeted sequencing” section. None of those references disclose the use of a single target-specific primer for target capture.

180. For example, Delseny describes methods of targeted sequencing for particular genes in rice species, and cites to Lu 2008 and Ammiraju 2008. EX1007.0417. Both of these studies utilize a bacterial artificial chromosome cloning system, wherein the genetic region of interest is introduced into a plasmid (a circular piece of DNA) that is then introduced into bacteria. EX2150.2072, EX2174.0001, EX2151.2305. As described in Lu 2008 and Ammiraju 2008, these plasmids are copied by the bacteria as the bacteria grow, after which they can be harvested and sequenced (in the case of Lu 2008 and Ammiraju 2008, through a non-discriminate shotgun sequencing method). This method of targeted sequencing does not use a single target-specific primer for target capture.

181. Delseny then discusses methods for “selecting relevant portions [of a genome that] might be of interest” and references a study describing an “exon capture” method designed to generate target specific circular DNA probes, Porreca 2007. EX1007.0417. Porreca 2007 discloses a method of generating circular “targeting oligonucleotide probes” (also known as padlock probes or molecular inversion probes) at scale. EX2128.0932. The targeting oligonucleotide probes are designed as nearly complete circular fragments with “targeting arms” (regions with sequences complementary to the sequence of interest) at the ends of the incomplete circle. *Id.* The targeting oligonucleotide probes are designed such that, when they are introduced to a sample with the genetic region of interest, each of the targeting arms will bind to that region. *Id.* The genetic region between the two arms can be enzymatically copied into the probe to complete the circle, which can then be closed using a ligase enzyme. *Id.* In this way, genetic regions of interest can be copied into the circular targeting oligonucleotide probes with sequence specificity, and the circularized targeting oligonucleotide probes can then be used for subsequent amplification via universal PCR and sequencing. *Id.* As with the prior bacterial artificial chromosome method, the target capture method referenced by Delseny does not use a single target-specific primer for target capture.

182. Delseny hypothesizes that methods such as target capture will be applied to “assess the genetic diversity associated with a number of genes

contributing to important agronomic traits,” explaining that this is one of the main goals of the international Generation Challenge Program.¹⁵ EX1007.0417. In support of this hypothesis, Delseny cites to Glaszmann, a review publication that discusses the potential of whole genome and targeted genome analysis methods to improve the understanding of crop genetics. *Id.*; EX2129.0167. Glaszmann particularly focuses on the importance of generating defined reference datasets to facilitate information sharing and comparison between researchers. EX2129.0171. Glaszmann does not provide particular information on any target capture methods, nor describe any method using a single target-specific primer for target capture.

183. To illustrate the benefits of comparing SNPs between plant lines, Delseny further cites to Wright. EX1007.0417. Wright “report[s] single-nucleotide polymorphism (SNP) diversity in 774 gene fragments (100 to 900 base pairs) in a sample of 14 maize inbred lines (representing modern maize) and 16 inbred teosintes (tables S1 and S2)” and conducts an extensive analysis of the variations in SNPs between the samples. EX2130.1310. In order to sequence the SNPs, Wright employed gene-specific primer pairs wherein “PCR primers [for a total of 1000

¹⁵ The Generation Challenge Program “was founded on the theme of ‘unlocking genetic resources for the resource-poor,’” such as developing drought-resistant crops. EX2112.

genes] were designed on the 3' UTR using Primer3.” EX2165.0001. The genes were amplified by PCR and sequenced using automated Sanger Sequencing. *Id.* Wright used the automated base calling methods PHRED and PHRAP to identify SNPs and set quality score thresholds, and reported that “[u]sing these criteria, we identified zero errors in 260,981 bp of sequence comparisons between replicates, suggesting our dataset has an error rate very close to zero ($< 3.8 \times 10^{-6}$ errors per bp) under these conditions.” EX2165.0001-02. The capture method described in Wright used two target specific primers for each gene of interest, and as such the method of targeted sequencing does not use a single target-specific primer for target capture. Importantly, it also reported a high accuracy rate for identifying SNPs using *pairs of target-specific primers*.

184. As described above, the only method of target capture that is expressly described by Delseny (rather than in a reference cited by Delseny) is the protein capture technique, chromatin immunoprecipitation. EX1007.0417-18. Delseny cites to a review on chromatin immunoprecipitation by Peter Park (Park 2009), wherein Park describes “the benefits and challenges in harnessing this technique with an emphasis on issues related to experimental design and data analysis.” EX2132.0669. Chromatin immunoprecipitation leverages the presence of proteins naturally bound to or associated with DNA in cells and selectively isolates those proteins (and associated DNA) from the rest of the cell. EX2132.0671. The DNA

can then be extracted from those proteins, further processed, and sequenced to determine the sequence of DNA associated with those proteins. *Id.* In this way, DNA-protein interactions and facets of DNA regulation can be explored. Park 2009.0669. Chromatin immunoprecipitation does not use any primers for targeted DNA capture (instead relying on different proteins), and as such, Park does not describe any method using a single target-specific primer for target capture. Delseny discloses that the chromatin immunoprecipitation has been used in maize (citing to Wolfgruber) and separately to “identify the target sequences of the Arabidopsis transcription factors [DNA associated proteins] SEPALLATA 3 and APETALA 1,” citing to Kaufmann 2009 and Kaufmann 2010. EX1007.0417. All of these articles describe chromatin immunoprecipitation capture methods, and none of them describe any method using a single target-specific primer for target capture. *See, e.g.,* EX2114, EX2115, EX2173.

185. Finally, Delseny describes an “application of targeted sequencing” wherein “[t]he sequences of one (or a limited number of) short DNA fragment(s) are used as a ‘barcode’ to identify and distinguish species from each others.” EX1007.0417. Delseny notes that while “this concept has rapidly progressed in the animal systematics community, debates within the botanical community concern the fragments to be chosen. However, recent progress indicates that a plant barcode might become a reality fairly soon.” *Id.* To support this point, Delseny cites to an

article from the CBOL Plant Working Group published in 2009 (“CBOL 2009”). *Id.* CBOL 2009 describes a pilot study testing the viability of “sequencing a standard region of DNA as a tool for species identification.” EX2135.12794. To test this, CBOL 2009 sequenced several different regions across different samples using “a single *primer pair* for each locus.” *Id.* (emphasis added) As such, the method of targeted sequencing described in CBOL 2009 and cited in Delseny does not use a single target-specific primer for target capture.

2. Delseny’s Discussion of Using NGS for SNP Genotyping Concerns Whole Genome Sequencing

186. In a section entitled “Population Structure, genetic diversity and genotyping,” Delseny explains that “[m]ajor challenges in plant breeding are the discovery of genetic variation within a species,” among other challenges. EX1007.0416. Delseny notes that “SNPs are the most abundant [genetic] variants.” *Id.* Delseny then discusses several approaches to genotyping.

187. Delseny first discusses the use of microarrays for genotyping, noting that SNP genotyping by Affymetrix DNA chips were “good enough to produce ~350,000 high-confidence SNPs with a false discovery rate of <2%.” *Id.*

188. Delseny then discusses several other approaches using microarrays. *Id.* Delseny notes that while the observations are important, “these observations give an incomplete view of genetic variation between many SNPs are missed and because InDels and repeats cannot be investigated by hybridization techniques.”

EX1007.0417. Delseny goes on to say that researchers “will soon be overwhelmed by new data resulting from NGSTs.” *Id.* A POSA would have understood Delseny to be referring to the known higher throughputs for NGS compared to microarrays.

189. For NGS genotyping, Delseny discusses a few re-sequencing studies. *Id.* A POSA would have understood Delseny to be referring to the practice of sequencing a sample that already has an existing reference genome when referring to “re-sequencing.” Delseny confirms as such by explaining that these re-sequencings are part of “a much larger project,” where re-sequencing occurs on a “small number” of reference genomes. *Id.* From these re-sequences, the authors reported that “67% of the aligned sequences are error-free,” meaning that one-third of aligned sequences were not (nearly 90% of the sequences could be aligned to the reference sequence). *Id.* After discussing the ongoing work of a few other projects, Delseny concludes that “[n]evertheless, “identifying high-confidence SNPs and small InDels remains a major challenge until base-calling is improved and false discovery rate is reduced.” *Id.*

190. In comparing microarrays to NGS, Delseny explains that “[s]o far, genotyping has involved mostly microsatellites, a technology with limited resolution and requiring much time and human resources.” *Id.* According to Delseny, “[t]he ultimate resolution is at the base-pair level and NGSTs now open the possibility to genotype a line by sequencing it, provided a reference genome is available.” *Id.* A

POSA would have understood that by “resolution,” Delseny is referring to the fact that microarrays cannot identify the nucleic acid sequences of an entire sample—microarrays rely on array probes that are specific to target sequences.

191. To highlight the concept of using NGS for genotyping, Delseny discusses a paper concerning the re-sequencing of rice with Illumina. *Id.* That paper is referenced in end-note 125, *High-throughput genotyping by whole-genome resequencing*. *Id.* It is the only paper referenced by Delseny as part of the disclosure that NGS can be used for genotyping. As discussed below, a POSA would have understood from this paper that when Delseny states “genotype a line by sequencing it,” Delseny is referring to genotyping an entire group of genetically similar plants. Delseny concludes the section by noting that “NGSTs have opened the way to molecular ecology because the very high throughput allows for investigating genomes at the individual and population levels.” *Id.* Again, as discussed in the next section, a POSA would have understood that these levels are referring to specific plant lines, whether an individual plant line or a population of plant lines. A POSA would also have understood from the disclosure that the genotyping through NGS discussed in Delseny concerns whole genome sequencing.

E. Han 2009 (EX2117)

192. The senior author of the paper, Bin Han, is also a co-author of Delseny. Compare *id.*, .0407 to EX2136.1068. As such, I will refer to this paper as Han 2009.

Han 2009 explains that high-throughput genotyping “was initially accomplished by adopting microarray technology, which detects single nucleotide polymorphisms (SNPs) through hybridizing genomic DNA to oligonucleotides spotted on gene chips.” EX2136.1068. Han 2009 goes on to explain that “[a]lthough the goal of high-throughput was achieved, serious limitations remain for the array-based method,” including that it was “laborious, time-consuming, and expensive to design, produce, and process microarrays suited for specific mapping populations.” *Id.* Han 2009 notes that the “advent of the next-generation sequencing technology holds the promise for a methodological leap forward in genotyping and genetic mapping” because of how it increases “sequencing throughput by several orders of magnitude” and “allow[s] simultaneously sequencing a large number of samples using a multiplexed sequencing strategy.” *Id.*

193. Han 2009 purports to “describe the first high-throughput genotyping method that uses SNPs detected by whole-genome resequencing.” *Id.* Han 2009 distinguishes the SNP data it obtains from “traditional genetic markers primarily in two aspects.” *Id.*, .1069. The paper explains that “it is often not the case that all members of a recombinant population can be scored at a given SNP site.” *Id.* Importantly, Han 2009 also explains that “*an individual SNP site is no longer a reliable marker or locus for genotyping due to several potential sources of sequence errors.*” *Id.* (emphasis added). To address these aspects that are specific

to the use of NGS for genotyping, Han “developed a new analytical framework, that is, a sliding window approach *for evaluating SNPs collectively rather than individually.*” *Id.* (emphasis added). Han 2009 applied this method to the whole genomes of 150 rice recombinant inbred lines. *Id.*

194. In discussing genotyping with this method, Han 2009 reported the “presence of minority SNPs” that were the “result of sequence errors.” *Id.*, .1070. Han 2009 explained that “[b]ecause of these noise SNPs, the genotype of the [rice recombinant inbred lines] could not be simply determined based on individual SNPs. That is, a SNP can no longer be treated as a mapping locus, as in the traditional way of using molecular markers.” *Id.* In discussing their results, Han 2009 explains that “[o]ur studies of the rice [recombinant inbred lines] demonstrated the advantages of this new genotyping method over the commonly used PCR-based approach.” *Id.*, .1073. Han 2009 explains prior to the whole genome sequencing method, they genotyped rice lines by PCR amplifying SNP markers, which “took three people more than one year of intensive work to complete.” *Id.* Their new method of whole genome sequencing generated more coverage within just two weeks. *Id.* Han 2009 also explains that whole genome sequencing led to “markedly improved mapping accuracy and resolution.” *Id.*

195. Han 2009 discloses that their “genotyping method requires the availability of genome sequences of the mapping parents,” which means that “the

quality of the genome sequences becomes an important issue.” *Id.* In fact, Han 2009 explains that “[w]hen the genome sequences of both parents are subjected to high error rates, e.g., both generated from genome resequencing, the accuracy of the genotype calling could drop below an acceptable level more quickly.” *Id.* A POSA would have understood this disclosure to mean that the genome sequencing method used in Han should not be used to obtain a *de novo* genome sequence for SNP genotyping given error rates. Han 2009 explains that “as sequencing technology improves, the sequence-based genotyping method, which has already looked promising, will continue to gain higher accuracy and efficiency.” *Id.* Han 2009 concludes by explaining that “[s]equencing coverage can be scaled up any time for the entire or a part of the mapping population.” *Id.* A POSA would have understood the mapping population to refer to the group of plant lines that are being mapped—meaning that even at the smallest unit of the mapping population, it would still entail an entire genome.

196. In the “Methods” section at the end of the paper, Han 2009 confirms that it did not target any specific sequences and instead ligated Illumina universal adapters to all fragments before PCR amplification. *Id.*

197. As explained above, the Han paper discusses using NGS for genotyping where the entire genome is sequenced and a reference genome exists for comparison. In comparison, Meyer 2009 (EX2118) is an example of a prior art publication where

a reference transcriptome did not exist. EX2118. In Meyer, the authors sequenced the entire transcriptome of a coral larval species with the 454 sequencer. EX2118.0001. Meyer disclosed using a software program to predict the number of potential SNPs that were obtained by sequencing the entire transcriptome. EX2118.0009, .0016-16. To validate whether this software program was accurate, Meyer used target-specific primer pairs to target 20 specific SNPs and sequenced the targeted sequences with Sanger sequencing. *Id.* From this validation, Meyer concluded that the program had correctly predicted approximately 70% of the SNPs in the sequencing data. *Id.* With these sequences now known and validated, Meyer suggested that a future application of the data would be to design target-specific microarray probes that could be used for genotyping. EX2118.0011.

198. The disclosures of Meyer reflect how a POSA at the time would have been cautious with inferring genotypical information from even whole genome/transcriptome sequencing where no reference was available. A POSA would have instead verified the accuracy of potential SNPs based on methods that were perceived to be more accurate, such as using target-specific primer pairs with Sanger sequencing, as Meyer did. And once the POSA could verify the sequence associated with specific SNPs, the POSA would consider using microarrays for future genotyping given the target-specificity and accuracy conferred by the array probes.

VIII. Neither Shapero Nor Delseny Disclose All Elements of the Independent Claims

199. For the '012 patent, Shapero fails to disclose or teach multiple claim elements of claim 1. Shapero is not a method for sequencing an enriched nucleic acid sequence of interest. Shapero does not disclose or teach at least the following limitations:

- a. “wherein products of the amplifying comprise a 3' end with sequence complementary to a sequence on a surface;”
- b. “annealing a strand of the products of the amplifying to the sequence on the surface using the 3' end with sequence complementary to the sequence on the surface;” and
- c. “sequencing the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.”

200. For the '108 patent, Shapero does not disclose or teach at least the following limitation: “sequencing the amplified products comprising the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.”

201. These claim elements are directed to the incorporation of universal primer sequences that are subsequently utilized for NGS. Shapero does not teach any methods for NGS, including methods incorporating the universal primers that are used for NGS. Instead, Shapero teaches the use of hybridization arrays, which do not require the use of a universal primer sequencing for sample binding, as

discussed above. A POSA would not have understood that Shapero's method using hybridization arrays could be used for NGS, which was not commercially available when Shapero published.

202. For the '012 patent, Delseny fails to disclose or teach multiple claim elements of claim 1. Specifically, Delseny does not disclose or teach at least the following limitations:

- a. "annealing one or more oligonucleotides in solution to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the partial duplex adaptor, wherein the one or more oligonucleotides comprise
 - i. a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and
 - ii. a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;"
- b. "extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the partial duplex adaptor with a polymerase, thereby generating one or more oligonucleotide extension products comprising
 - i. sequence complementary to the first adaptor sequence at a first end,

- ii. sequence complementary to the nucleic acid sequence of interest, and
 - iii. the second adaptor sequence at a second end;” and
- c. “amplifying the one or more oligonucleotide extension products using
- i. a first primer that anneals to a complement of the first adaptor sequence and
 - ii. a second primer that anneals at its 3’ end to a complement of the second adaptor sequence to enrich for the nucleic acid sequence of interest,
 - iii. wherein products of the amplifying comprise a 3’ end with sequence complementary to a sequence on a surface.”

203. For the ’108 patent, Delseny fails to disclose or teach multiple claim elements of claim 1. Specifically, Delseny does not disclose or teach at least the following limitations:

- a. “annealing one or more oligonucleotides in solution in a reaction mixture to the nucleic acid sequence of interest in a nucleic acid fragment, wherein the reaction mixture comprises a plurality of nucleic acid fragments, wherein the nucleic acid fragment comprising the nucleic acid sequence of interest comprises a first adaptor sequence, wherein the one or more oligonucleotides comprise

- i. a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and
 - ii. a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;”
- b. “extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment comprising the first adaptor sequence with a polymerase, in the reaction mixture, thereby generating one or more oligonucleotide extension products comprising
 - i. sequence complementary to the first adaptor sequence at a first end,
 - ii. sequence complementary to the nucleic acid sequence of interest, and
 - iii. the second adaptor sequence at a second end;” and
- c. “amplifying the one or more oligonucleotide extension products, in the reaction mixture, using
 - i. a first primer that anneals to the complement of the first adaptor sequence and
 - ii. a second primer that anneals at its 3' end to a complement of the second adaptor sequence,
 - iii. thereby enriching the nucleic acid sequence of interest, by generating amplified products comprising the enriched nucleic acid sequence of interest.”

204. Delseny fails to teach these limitations because Delseny fails to teach the claimed methods for enriching sequences of interest through amplification, wherein those amplified products are subsequently sequenced through NGS.

205. The Petitions and Dr. Metzker's accompanying declarations do not dispute that Shapero fails to teach the use of NGS and that Delseny fails to teach a method of single primer enrichment through amplification.

IX. A POSA Would Not Have Been Motivated to Combine Shapero with Delseny With a Reasonable Expectation of Success

206. QIAGEN argues that a POSA would have been motivated to combine Shapero and Delseny, and in doing so, modify the MARA method by discarding its microarrays and using it with NGS. QIAGEN also argues that a POSA would have had a reasonable expectation of success in doing so. For the reasons explained herein, I disagree with both arguments.

A. A POSA Would Not Have Been Motivated to Adapt a Method Used with Hybridization Arrays for Use With NGS In the Manner Claimed by the Patents

207. Shapero discloses the use of the MARA method with a predecessor Affymetrix hybridization array for the genotyping of SNPs. *See supra*, ¶ 163. A POSA would not have been motivated to combine a publication on hybridization arrays with NGS technology to practice the claimed methods at least because the

technologies are fundamentally different, posing different challenges and making it difficult to extrapolate results from one to another.

208. As explained above, a hybridization array utilizes a silica chip containing hundreds of billions of custom probes that are designed to target specific sequences. *See supra*, ¶¶ 65-67. Generally, these arrays contain numerous probes specific to one target where the various probes are configured slightly differently to optimize their ability to capture the specific target sequence. *See supra*, ¶¶ 65-68. When samples are ran over the array, these probes bind their target sequences on the array. *See supra*, ¶¶ 64, 68, 111. Non-target sequences do not have a corresponding probe containing a complementary sequence, so those sequences are removed from the array during subsequent washing steps. *Id.*

209. By design, the hybridization arrays drive target recognition/specificity. *Id.* It is not necessary for a researcher to prepare a library that is enriched for target genomic material during sample preparation because the targeted design of the array. *See supra*, ¶¶ 64, 68, 111-12. Instead, the main concern during sample preparation for hybridization arrays is ensuring that enough material is being used to cover the array with an evenly distributed amount of material so as to ensure that target sequences come into contact with their corresponding array probe and are not inadvertently washed away. *See supra*, ¶ 68. Because of this concern for adequate sample material, researchers generally focus on non-specific amplification during

sample preparation as a means of ensuring an adequate amount of sample materials for the hybridization array. *See supra*, ¶¶ 67-68.

210. Due to the target specificity of the arrays and the washing steps that are used before array analysis, concerns with off-target binding during sample preparation are minimized. *See supra*, ¶ 112. Hybridization arrays are designed to filter out sequences that are not specific to the array probes. *See supra*, ¶¶ 64-66, 68, 111. After off-target sequences are eliminated, the arrays are prepared for image analysis. *See supra*, ¶¶ 45, 111. From there, researchers can run analyses to obtain outputs that show whether the specific target was present in the sample. *Id.*

211. NGS platforms, on the other hand, are designed differently. Rather than utilizing hundreds of billions of target-specific probes for analysis, NGS platforms seek to leverage a handful of universal primers. *See supra*, ¶¶ 106-107. These platforms are designed so that any sample fragment containing the universal primer sequence will be affixed to a surface—such as the surface of the sequencer—and then subsequently sequenced. *See supra*, ¶¶ 106-110. The use of universal primers is a main feature of NGS platforms. These universal primers allow researchers to sequence hundreds of thousands of different sample fragments in one sequencing run without the burden and expense of designing target specific primers. *See supra*, ¶¶ 81-88, 107-110. The ability to sequence so many different fragments at once

allows for the high-throughput sequencing that is characteristic of NGS. *See supra*, ¶¶ 81-88.

212. Because NGS uses universal primers for high-throughput sequencing, sample preparation techniques must be used to preserve target specificity and eliminate off-target binding. If the sample is amplified without any regard for enriching specific sequences, then all amplified products that contain the universal primer sequences will be sequenced. *See supra*, ¶¶106-110. This would be expensive and would generate a massive amount of data. To reduce this burden and focus the analysis, the researcher can enrich for (or isolate) specific sequence(s) of interest prior to sequencing. *See supra*, ¶¶87-89.

213. Above, I illustrated the critical differences in target recognition and capture between the analytical platforms. *See supra*, ¶¶112-113. The hybridization array probes will likely detect the presence of the targeted SNP without the need for any additional sample preparation considerations. *See supra*, ¶111. If the researcher attempts to use the same sample preparation methods for NGS—with the one exception being the additional step of incorporating the universal NGS primers, the data will include sample reads that reflect the sequence of the target SNP as well as sequencing reads for all the other sequences in the sample. *See supra*, ¶106-110. The considerably higher data output, along with read fidelity issues experienced by NGS, will pose significant challenges for reliably genotyping the SNP because of

the sheer volume of additional genomic information as well as the increased risk that any detection of the SNP in the sequencing data is a false positive attributable instead to read errors. *See supra*, ¶86.

214. Given these technology differences and data outputs, a POSA would have been unlikely to consider hybridization array techniques for use with NGS. A POSA would have known that due to the different issues and limitations with hybridization arrays, sample preparation techniques for hybridization arrays would not have been designed specifically for the purpose of promoting target specificity and minimizing off-target binding—issues heavily exacerbated with NGS.

215. Additionally, a POSA would not have been motivated to start with a technique specific to a hybridization array (even methods directed to the use of a single target specific primer for preparing libraries for amplification) when designing an NGS method given that the hybridization arrays would make it difficult, if not impossible, to infer how the sample preparation methods affected targeted specificity and off-target binding. The target specificity of the hybridization arrays masks any potential target specificity issues with sample preparation methods so that a POSA could not tell whether the sample preparation contributed to target specificity or diminished such specificity. *See supra*, ¶¶ 111-112. With the hybridization array, the array specificity coupled with washing steps also masks concerns of off-target binding so that a POSA could not tell whether the sample

preparation increased or decreased off-target binding. *See supra*, ¶¶ 64, 68, 111. As such, a POSA reviewing sample preparation methods for hybridization arrays (even those using a single, locus-specific primer) would be confronted with methods where target specificity and off-target binding were not primary concerns and where potential issues with such were masked by the use of the hybridization arrays. A POSA would be reluctant to use a sample preparation method for hybridization arrays as a starting point for NGS.

B. Shapero Fails to Resolve the Concerns with Target Specificity and Off-Target Binding

216. There is nothing different about the Shapero publication that would make it more likely that a POSA would be motivated to adapt a sample preparation method for hybridization arrays so that it could instead be used for NGS. In fact, Shapero exemplifies the reasons why a POSA would not have been motivated to adapt a method like MARA for NGS, including because (1) Shapero discloses that target specificity is achieved by the hybridization arrays; (2) the experimental data reported in Shapero is only specific to hybridization arrays; and (3) the experimental design used to obtain that data makes it impossible for a POSA to infer what impact, if any, MARA had on target specificity and off-target binding.

217. First, Shapero tells us that target specificity is achieved through hybridization arrays. In discussing the MARA protocol, Shapero notes that “[i]mportantly, MARA does not require absolute specificity of primer annealing and

elongation since target recognition and allele discrimination is achieved through hybridization to high-density arrays.” EX1006.0003-04. A POSA reading this statement would have understood that, like with other array hybridization methods, Shapero is relying on the target-specific array probes to recognize SNPs. A POSA would also have understood that target recognition and allele discrimination did not occur as part of sample preparation.

218. Shapero notes that “[w]ith MARA, allele discrimination is achieved by hybridization to arrays containing SNP-specific complementary sequences rather than a universal array which requires the use of tags or zip codes attached to the primers for genotype de-convolution.” EX1006.0007. “Thus, the requirement for absolute specificity of primer annealing to targeted sequences in genomic DNA is alleviated.” *Id.* A POSA reading this statement would again have understood that target specificity is being driven by the hybridization arrays, not the sample preparation. For example, “allele discrimination” is synonymous with target specificity. It is the concept of identifying specific alleles from a set of alleles. Shapero explicitly teaches that this discrimination “is achieved by hybridization to arrays” that contain target-specific probes. A POSA would have understood this statement to also mean that the locus-specific primers of the MARA method are not what is achieving hybridization.

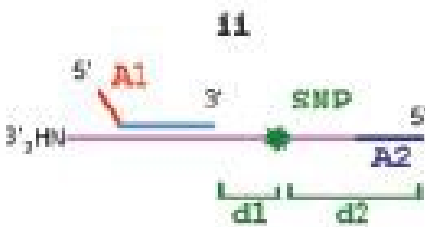
219. Shapero further reinforces the reliance on the hybridization array to promote target specificity in discussing off-target binding. Shapero notes “even if the primers contribute to non-specific amplification products, the complexity should not reach a level that would affect array hybridization.” *Id.* A POSA reading this statement would have understood that Shapero was relying on the hybridization array to ensure that off-target products did not impact the analysis. This is because the hybridization arrays used in Shapero were designed to handle an even greater capacity of genetic material than what Shapero tested. EX1006.0007 (“In the case of the 4365-plex MARA, this is ~2 MB. The high-density genotyping arrays are capable of handling DNA targets that are ~300 Mb of sequence complexity.”). A POSA would have expected that so long as the capacity was not exceeded, the microarray should be able to filter off-target products. That is unlike with NGS where there is no target-specificity conferred by the platforms to enable the filtering of off-target products. A POSA considering Shapero’s statements concerning target specificity and off-target binding would have concluded that Shapero relied on the hybridization array and offered no disclosures as to how the MARA method impacted either.

220. Second, QIAGEN argues that the data of Shapero would have motivated a POSA to adapt the MARA method for use with NGS. Paper 2, 34-48. That is incorrect. QIAGEN specifically relies on two sets of data, comparing the

call rates where locus-specific primers were used for the MARA method and where the primers were not. As explained above, Shapero relies on the call rates coupled with discrimination scores (“DS”) to evaluate the percentage of SNPs that were detected by the hybridization arrays. *See supra*, ¶¶ 162-64. The data is wholly obtained through the use of hybridization arrays and not NGS platforms. *See supra*, ¶¶ 150, 162-64. Given the fact that target specificity was driven by the hybridization arrays, a POSA could not infer from the data what degree of target specificity was driven by the MARA method, let alone how that method might translate to use with NGS where there are not safeguards to ensure target specificity.

221. Third, QIAGEN claims specifically the difference in call rates for the MARA 250-plex and MARA 750-plex demonstrate that the MARA method could be used with NGS given its high degree of target specificity. Paper 2, 41-42. QIAGEN, however, misreads the data reported by Shapero. For both sets of multiplex reactions, Shapero examined what the call rates and discrimination scores were for two different sample preparation methodologies. *See supra*, ¶¶ 162-65.

222. As explained above, the first methodology relied on the MARA method is illustrated in Figure 1 of Shapero, where one of two amplification primers (A1) is incorporated through the use of a locus-specific primer. *See supra*, ¶¶ 153-59.



EX1006.0004. As Shapero explains, a primary set of PCR reactions occur where the locus-specific primers are extended to incorporate the complement of the A2 sequence. Then, the MARA method relies on “cycles of secondary PCR with universal A1 and A2 primers [to] amplify all full-length products derived from the first round locus-specific reaction.” *Id.* These amplified products are then prepared for array hybridization. *Id.*

223. The second methodology omits the locus-specific primers in the first reaction. *See supra*, ¶¶ 164-65. Because of that, there is no incorporation of amplification primer A1. *Id.* As a consequence, there are no locus-specific extension products and the fragments do not undergo amplification as described in the standard MARA method. *Id.* Shapero acknowledges this by stating that the “results show that the presence of locus-specific primers allows specific target sequences to be amplified, and in the absence of primers, only small quantities of non-specific products are generated.” EX1006.0005. A POSA would have understood that “non-specific products” in this context refers to the fragment population as a whole (as opposed to the off-target products that were generated through MARA).

224. This distinction between the two methodologies is critical because it directly impacts what can be inferred from the Shapero data. Shapero is comparing instances where fragments are amplified to instances where fragments are not. As such, a POSA would understand the differences in call rates and discrimination scores to largely be impacted by the amount of genetic material used with the hybridization array. Where the genetic material was amplified before analysis, Shapero obtained better coverage and better results. Where the genetic material was not amplified, Shapero saw worse coverage and worse results.

225. A POSA would have understood that microarrays require a minimum amount of genetic material to adequately cover the microarray surface and expose the probes to the genetic material. Early microarrays required large amounts of genetic material to allow for reliable sample detection. By the mid-2000s, improvements in microarray workflows allowed many arrays to function reliably with ~1 microgram of sample. EX2138.0001. Efforts to further reduce the amount of required starting material focused on using PCR amplification to increase the amount of starting material prior to loading the sample onto the microarray. *Id.*, see also EX2139.0997. Using PCR methods, researchers were able to generate ~6-12 micrograms of DNA from 5-100 nanograms of starting material. EX2138.0003. That a minimum amount of material was needed for adequate microarray function was a known feature of microarrays, and researchers often went well above this

minimum when using microarrays to ensure robust signal intensity. For example, Matsuzaki *et al.* used ~30 micrograms of DNA for a 8- μ m array designed to genotype ~50,000 SNP. EX2167. Indeed, as late as 2009 Affymetrix microarrays “required” 500 nanograms of DNA for use. EX2141.0003. A POSA would have understood that Shapero’s omission of the locus-specific primers would have likely resulted in too little genetic material to ensure adequate array coverage.

226. Shapero also does not disclose that the data shows high levels of target specificity by the MARA method or even that the MARA method promoted high levels of enrichment for specific loci. That is because the experimental design described above precludes such conclusions. A POSA cannot infer that the MARA method promoted target specificity given the use of hybridization arrays in both instances (which Shapero tells us drives target specificity) and that Shapero did not amplify the fragments in one instance. Had Shapero amplified the fragments in both instances using amplification primer A1, then it is possible that a POSA could infer that differences in the call rates and discrimination scores are attributable to some level of enrichment. But that is not the case here. Additionally, given the same parameters, a POSA could not conclude with confidence to what degree the MARA method actually enriched any of the target loci. This is, in part, because of the inability to measure the degree of off-target binding. Shapero fails to disclose the relative populations of various SNPs pre- and post-MARA. Shapero also fails to

provide a reference amplification population by which a POSA could compare the post-MARA populations of target sequences to assess enrichment. Without these disclosures, a POSA cannot reliably infer to what degree MARA enriched the sample.

227. Further complicating the data are the concordance rates reported by Shapero, which showed that the uses of the hybridization arrays in Shapero created approximately four-to-five times as many errors (discordances) as the standard uses of the same type of hybridization array. *See supra*, ¶¶ 162-68. As Shapero explains, the use of MARA with the p502 microarray was compared to using the same microarray with the “whole genome sampling assay” that was described in Kennedy. EX1006.0003. Kennedy (EX2035) describes a method in which total genomic DNA is digested with restriction endonucleases, adapters are ligated to both ends of the fragments, and primers specific to the adapters are used to PCR amplify the fragments (preferentially selecting for 400- to 800-bp fragments). EX2035.1234. After PCR amplification, the amplicons are further fragmented, labeled, and hybridized to a special microarray that is optimized to bind to specific SNPs. *Id.*, .1233-34. Kennedy discloses that the “study represents a proof of principle for the simultaneous genotyping of 14,548 SNPs without locus-specific PCR.” *Id.*, .1235. In other words, the method of Kennedy (WGSa) is designed to intentionally omit the locus-specific primers used in Shapero.

228. As Shapero explains, the concordance rates are meant to compare the “MARA genotype calls with WGS genotype calls in order to determine the concordancy of MARA calls with the alternative WGS target preparation.” EX1006.0005. And as I explain above, when comparing MARA to the WGS method, Shapero reported four-to-five time higher error rates. *See supra*, ¶¶ 164-68. To the extent a POSA could make any inferences based on the data reported in Shapero, it would be that the WGS method was more accurate than MARA. As such, these heightened error rates would only further diminish any interest a POSA would have in using the reported method for new genotyping methods, given that the reported methods were less accurate than other known uses of the hybridization arrays, and it would certainly reinforce why a POSA would not be motivated to adapt this method for NGS, where it was very likely that the genotyping error rates would only increase. *See supra*, ¶¶ 86, 105-10, 175, 186-90.

229. Putting all of this together, a POSA would not have been motivated by Shapero to modify the MARA method for use with NGS. A POSA would not have started with Shapero given that the MARA method was not designed for NGS, not tested with pre-NGS sequencers, and instead relied on hybridization arrays to ensure target specificity. The data in Shapero was obtained through using the hybridization arrays, and the experimental design would have precluded a POSA from being able

to infer the impact of the MARA method on enrichment, target specificity, and off-target binding.

C. A POSA Reviewing Shapero Would Not Have Discerned a Need to Switch from Hybridization Arrays to NGS

230. If a POSA were looking at Shapero for genotyping, a POSA would not have thought that the MARA method could be improved by using it with another analytical platform, whether pre-NGS Sanger sequencing or NGS. Rather, the art at the time promoted the efficacy of using hybridization arrays for the kind of SNP genotyping performed in Shapero. *See supra*, ¶¶ 65-67, 74. The main issue in the art at the time was how to increase high-throughput genotyping with hybridization arrays. *See supra*, ¶¶ 65-67, 74, 151. As discussed further herein, a POSA would have perceived this limitation to also impact the hybridization arrays disclosed in Shapero given the unfavorable results observed by Shapero with the higher multiplex reactions. *See supra*, ¶¶ 167-69. If anything, a POSA would have been motivated to improve on the hybridization array hardware or software. *See* EX2004.0024 (“Since their inception, high-density microarrays have followed the same trend as computer microprocessors. In 1965, Moore predicted that the power of microprocessors would double every 18 months. This trend has been true for the computer industry, where much faster and smaller central processing units are

constantly being produced. High-density oligonucleotide arrays have evolved similarly.”).

231. For example, Affymetrix continued to develop its microarray technology to the point that by 2008, Affymetrix arrays were recognized as “widely used across all fields of biological research for measuring genome-wide gene expression.” EX2142.0001. Indeed, the Affymetrix GeneChip microarrays were considered by some researchers as “the most popular high density oligonucleotide gene expression arrays and [] an invaluable tool in genomics studies worldwide.” EX2142.0002. This popularity could be partially ascribed to “key differentiators between Affymetrix’s GeneChip technology and other microarray-based methods” such as the manufacture of Affymetrix arrays by photochemical synthesis, which allowed for “more than a million different probes [to] be synthesized on an array roughly the size of a thumbnail.” EX2143.0003. This scale allowed for both increased throughput and “the inclusion of multiple probes to interrogate the same target sequence, providing statistical rigor to data interpretation.” *Id*, see also EX2144.0001.

232. Throughout the 2000s, Affymetrix genotyping arrays increasingly focused on increasing genomic SNP coverage, adding more SNP probes to the arrays. EX2144.0001, EX2141.0001. Genome-wide SNP analysis was noted for “not requir[ing] site-specific primers, [being] highly scalable, and enabl[ing] the

creation of hybridization target starting with as little as 250 ng of chromosomal DNA.” EX2143.0015. Such platforms included the GeneChip® Human Genome U133 Arrays, which employed a multi-array plate format that allowed for parallel analysis of multiple whole genomes while “reducing labor cost and increasing standardization with the genome.” EX2144.0001. Another example was the Genome-Wide Human SNP Array 6.0, which “feature[d] more than 1.8 million markers of genetic variation” and “allow[ed] researchers to perform association studies with large sample sizes in both initial scan and replication phases, thereby significantly increasing the overall genetic power of their studies.” EX2141.0001. The Genome-Wide Human SNP Array 6.0 was described as “demonstrat[ing] industry-leading performance and represent[ing] more genetic variation on a single array than any other product, providing maximum panel power and the highest physical coverage of the genome.” *Id.* And in fact, the Genome-Wide Human SNP Array 6.0 is still used today for molecular cytogenetics. As such, a POSA would not have been motivated to decouple Shapero from its microarrays. If anything, a POSA would have been motivated to improve upon the microarrays. Even if a POSA would have been motivated to apply the teachings of Shapero for use with NGS (which a POSA would not have been motivated to do), a POSA would not have been motivated to ditch Shapero’s microarrays, especially where various hybrid capture

methods were known in the art and had been reported on favorably. *See infra*, ¶¶230-33, 255-260.

233. QIAGEN's assumption that a POSA would have automatically viewed the way to improve on Shapero as integrating other technologies (such as NGS) reflects nothing more than hindsight.

D. A POSA Would Not Have Understood Shapero's Reference to Syvänen to Teach that the MARA Method Could be Used with Sequencing Technologies

234. QIAGEN argues that one reference to Syvänen by Shapero would have been understood by a POSA to mean that Shapero was disclosing that the MARA method could be used with sequencing technologies that existed at the time. Paper 2, 34-35. I disagree. In the introductory sections of the publication, Shapero refers in passing to the "multitude of genotyping methodologies for SNPs that have been described over the years." EX1006.0001. Shapero proceeds to note that "[t]hese approaches can vary in terms of how alleles are discriminated, e.g. the use of nucleic acid hybridization with short oligonucleotide probes alone or in combination with enzymes such as DNA polymerases or DNA ligases" and "in the nature of both the assay formats and labeling and detection strategies." *Id.* Shapero proceeds to state that "many currently used approaches share a common feature in that there is a requirement for an amplification of the region harboring the polymorphism prior to actual genotyping." *Id.* At no point does Shapero state that these "currently used

approaches,” especially as reflected in Syvänen, show that the MARA method could be used with sequencing technologies. In fact, I have not seen any disclosure to that effect in any part of Shapero.

235. QIAGEN argues that Syvänen’s disclosures concerning the use of DNA-sequencing instruments with “minisequencing” support that the MARA method could have been used with sequencing technologies. This argument misunderstands what Syvänen discloses. First, “minisequencing” is a term used by Syvänen to refer to sample preparation methods in which an extension product is created that allows for the subsequent detection of a SNP. *See supra*, ¶¶171. It does not by itself refer to sequencing technologies such as Sanger and NGS. *See supra*, ¶¶170-71. Second, Syvänen discusses the use of DNA-sequencing instruments in the context of how certain features of the instruments could be used for SNP detection. *See supra*, ¶172. As an example, Syvänen discusses that 96-channel capillary sequencers (i.e. pre-NGS Sanger sequencers) could be used to take advantage of the electrophoretic size-separation features of the sequencers. *Id.* Syvänen proposes using size separation as a means of high-throughput genotyping, not comparing Sanger sequencing read outputs for genotyping. *Id.* As an example of the point, Syvänen cites to SNaPshot, which as I explained above, was an automated system that analyzed a single fluorescent nucleotide during extension. *Id.*

A POSA would not have understood that these disclosures supported the use of the MARA method with sequencers.

236. In fact, Syvänen would have discouraged the adaptation of MARA for SNP genotyping with sequencing technologies. As discussed above, Syvänen discusses the use of allele-specific primers in PCR. *See supra*, ¶171. Syvänen discloses that by using such primers “without a consecutive target-specific detection reaction or size-separation step, the specificity of the method might be hampered owing to primer dimers and other spurious amplification products that will not be distinguished from actual PCR products.” EX1012.0935. In fact, Syvänen notes that a “limitation of all variants of allele-specific PCR is that the reaction conditions or primer design for selective allele amplification must be optimized empirically for each SNP,” and as such, “are best suited for the analysis of a limited number of SNPs in large sample collections.” *Id.* A POSA reading these disclosures would understand that Syvänen was advocating for methods like MARA to be used with hybridization arrays (which can serve as a consecutive target-specific detection reaction) and that such sample preparation methods would be ill-suited for high-throughput sequencing, such as NGS. Further, while Syvänen makes a similar observation as Shapero that many genotyping methods relied on PCR amplification, Syvänen discourages the use of PCR amplification in the “next generation technology for SNP analysis.” EX1012.0940. Rather, Syvänen discloses that the

technology “should avoid the PCR amplification step, and the technology could be based on the analysis of single DNA molecules, which would allow direct determination of the haplotypes formed by the SNPs.” *Id.* In a table where Syvänen identifies the “most significant advantage,” “disadvantage,” or “special feature” of various SNP-genotyping methods, Syvänen recognizes that the most significant advantage of one technology is that “PCR amplification [was] avoided” and that a special feature of another is that it “avoids target amplification.” EX1012.0934. Syvänen’s concerns with PCR amplification culminate in Syvänen remarking, “[h]opefully, future SNP-genotyping technology will be more elegant than the incremental and brute-force expansion of existing technologies in the way that the Sanger sequencing method has been exploited for genome sequencing.” EX1012.0940. In view of Syvänen, a POSA would not have been motivated to modify the MARA method, which utilizes PCR amplification, for genotyping with sequencing technologies. Instead, a POSA would have been discouraged from considering the MARA method for SNP genotyping with newer technologies, like NGS, given these disclosures by Syvänen.

E. A POSA Would Not Have Been Motivated by Delseny to Adapt the MARA Method for NGS

237. QIAGEN argues that Delseny would have given a POSA the motivation to adapt the MARA method for use with NGS given that “Delseny recognizes that NGS is a replacement for genotyping described in Shapero.” Paper 2, 35-36. I

disagree that Delseny taught that NGS is a replacement for high-throughput allele discrimination for known SNPs, which is the genotyping described in Shapero. As I explain below, QIAGEN's interpretation of Delseny is overbroad. Delseny is advocating for, at best, the potential use of NGS for genotyping where the whole genome is sequenced and can be compared to a reference genome (reflecting concerns even then with accuracy). *See supra*, ¶¶ 186-98. For targeted capture (unlike whole genome sequencing), Delseny would have encouraged the use of pre-NGS Sanger sequencing for genotyping.

238. QIAGEN relies on three specific disclosures from Delseny to argue that a POSA would have been motivated to combine Shapero with Delseny: (1) “NGS ‘allow[s] for many more samples to be analyzed at the same time,’” (2) “So far, genotyping has involved mostly microsatellites, a technology with limited resolution and requiring much time and human resources. The ultimate resolution is at the base-pair level and NGSTs now open the possibility to genotype a line by sequencing it, provided a reference genome is available;” and (3) “Such an approach can substitute for the genotyping microarray strategy because of its higher resolution.” *See Paper 2*, 35-36. A POSA would have understood that these disclosures reflected the thinking of the time that high-throughput methods, such as NGS, allowed for researchers to more efficiently than ever before obtain large amounts of genomic information. As Delseny discloses, researchers could sequence more samples at

once. Additionally, rather than focusing on specific genotypes, researchers could process reams of sequences to assess numerous genotypes. As such, Delseny recognized that the “higher resolution” provided by NGS “can substitute for the genotyping microarray strategy.” EX1007.0417. Delseny does not teach, however, that NGS should be a substitute for the kind of genotyping used in Shapero, where Shapero was seeking to optimize a hybridization array that would allow for the screening of known SNPs in genomic material. *See* EX1006.0001.

239. Critical to how a POSA would understand Delseny’s disclosures concerning using NGS for SNP genotyping is what Delseny is citing. As I discuss above, Delseny cites one reference for the proposition: Han 2009. *See supra*, ¶ 191. Han 2009 concerns the re-sequencing of the entire genome of rice recombinant inbred lines. *Id.* It does not employ any form of target capture. *Id.* And in fact, it discourages the use of target capture in a few ways.

240. First, it explicitly warns that “an individual SNP site is no longer a reliable marker or locus for genotyping due to several potential sources of sequence errors” caused by NGS. *Id.* It explains that “[b]ecause of these noise SNPs, the genotype of the [rice inbred lines] could not be simply determined based on individual SNPs. That is, a SNP can no longer be treated as a mapping locus, as in the traditional way of using molecular markers.” *Id.* A POSA would have

understood that to mean that MARA should **not** be used with NGS because MARA relies on locus-specific markers.

241. Second, it teaches that “[t]o deal with these unique features of the SNP data generated by the next-generation sequencing, [they] developed a new analytical framework, that is, a sliding window approach for evaluating SNPs collectively rather than individually.” *Id.*

242. As I explained above, a POSA would have reviewed the disclosure cited by Delseny to understand Delseny’s commentary. From reviewing Han 2009, it is my opinion that my fellow POSA would have understood the specific disclosures relied on by QIAGEN at EX1007.0417 to mean the following:

- a. “The ultimate resolution is at the base-pair level”: *It would be optimal to be able to genotype SNPs by reviewing data that identifies the specific base pairs of the sample.*
- b. “and NGSTs now open the possibility to genotype a line by sequencing it, provided a reference genome is available”: *Because NGS allows for the sequencing of an entire genome for a group of genetically related plants (the plant line), it could be used to genotype those plants; however, given the known error rates with NGS, the re-sequenced genome would need to be compared to a reliable reference genome.*

- c. “Such an approach can substitute for the genotyping microarray strategy because of its higher resolution”: *Whole genome sequencing that avoids locus-specific markers could replace genotyping microarrays since NGS can provide more sequencing data.*
- d. “NGSTs have opened the way to molecular ecology because the very high throughput allows for investigating genomes at the individual and population levels”: *NGS is able to sequencing a higher multiplex of samples, which allows for whole genome sequencing of a single plant, plant line, or group of plant lines.*

243. None of these disclosures teach the use of target capture with NGS for SNP genotyping. When considered in context, these disclosures teach away from that very occurrence.

244. Delseny also discloses that early NGS technologies suffered from a series of inaccuracies, which would have diminished their ability to be used for SNP genotyping. *See supra*, ¶¶ 186-90; EX1007.0417 (“Nevertheless, identifying high-confidence SNPs and small InDels remains a major challenge until base-calling is improved and false discovery rate is reduced.”). For example, Delseny discloses

that NGS platforms suffered from base calling issues,¹⁶ which could contribute to read artifacts and false positives. *See supra*, ¶¶ 175, 186-90. To illustrate these deficiencies, Delseny explains how 454 sequencers struggled to sequence homopolymer sequences because of how stretches of the same repeating nucleotide would throw off the calibration of the machine, leading to read artifacts or incorrect sequencing reads. *See supra*, ¶ 175. A POSA would have understood that read fidelity issues such as these would pose particular concerns with SNP genotyping where the purpose of the research would be to understand how a single nucleotide varies between samples. Given these read fidelity issues of NGS, Delseny reports that “[s]o far, the Sanger method has the highest accuracy because [of] *E. coli* DNA-polymerase and remains the ‘gold standard.’” EX1007.0411.

245. A POSA considering Delseny in its entirety would not be motivated to modify the MARA method for NGS. The POSA would understand that certain genotyping could be accomplished through NGS and that in some instances, that genotyping could replace the use of hybridization arrays. But the POSA would have understood those instances required whole genome sequencing and then comparison to a reference genome, as explained by Delseny and Han. This is because the read

¹⁶ Base-calling refers to the process by which a sequencer reads the specific nucleotides in a nucleic acid sequence.

fidelity issues with NGS made it less reliable at SNP genotyping through allele discrimination than other methods, such as Sanger sequencing and array hybridization.

F. A POSA Would Not Have Been Motivated to Modify the MARA Method to Practice NGS Given the Weight of the Prior Art

246. As I explain above, the POSA would have been aware of decades of publications criticizing the accuracy of a single target-specific primer, especially with PCR amplification, given issues with off-target binding. *See supra*, ¶¶ 57-62.

247. For example, some of the earliest target-specific primer genome walking methods were developed expressly to address issues with “[i]llegitimate products result[ing] from non-specific annealing of either primer,” which was characteristic of the earliest vectorette PCR reactions. EX2044.1644.

248. Siebert recognized these issues as early as 1996. Siebert noted that many single target-specific primer genome walking methods suffered from “a high background of non-specific PCR products,” which often made “it necessary to do a second round of PCR with nested primers to achieve greater specificity.” EX2043.0526. Siebert explained that using PCR suppression could improve on the technique; however, Siebert also acknowledged that the amplification of some cDNA still yielded high levels of unwanted amplification products. *Id.* In those instances, Siebert recommended the use of additional, more complex methods to filter out background products. *Id.*, .0006.

249. Similar observations appear in the art only a few years later in 1998, when Padegimas *et al.* attempted to address the “low specificity, low efficiency, and/or extensive cycling that contributed to a higher error rate in PCR products” in then-current PCR based walking methods, including vectorette, splinkerette, and adapter-based methods. EX2045.0149. Notably, the single target-specific primer method advanced by Padegimas required both a non-target degradation system and extensive optimization to achieve a reduction in off-target product generation. *Id.*, .0003-05.

250. Despite these efforts, more than seven years later Miura *et al.* disclosed that “single unique primer in combination with a common primer” genome walking methods were understood to “often [be] plagued with non-specific amplification products derived from so-called ‘off-targets.’” EX2047.4363. Further, Miura understood that this off-target phenomena was characteristic of single target-specific primer techniques rather than conventional PCR amplification methods, which “use[] a pair of unique primers to selectively amplify the target sequence” that achieve specificity “by a synergistic effect between the specificities of the two primers.” *Id.*

251. Nearly five years later, Bowers *et al.* noted that “[d]espite claims of specific amplification of the desired [] region of a target gene, a common problem that occurs in RACE reactions is obtaining non-specific amplification products.”

EX2049.0002. Similarly, a year later, in a patent filing, Sohn *et al.* reiterated that while “many studies have been made to improve the genome walking method,” “it is still in need of improvement.” EX2050.0002. Sohn stated that the “most serious problems of the PCR-based genome walking methods are low amplification efficiency and non-specific amplification.” *Id.* Sohn explained that to “improve this problem, complicated multi-step methods were developed, including secondary PCR, purification of desired PCR products using biotinylated primers, or improvement of PCR cycle conditions.” *Id.* Even then, “satisfactory performance [had] not yet been achieved.” *Id.*

252. More broadly, the art taught away from and expressed skepticism at various amplification methods. Tsoktouridis *et al.* conducted a comparison of various PCR amplification methods and found “no simplicity, low reproducibility, high backgrounds of nonspecific PCR products, insufficient homopolymer tailing, small PCR product amplification, and laborious time-consuming procedures.” EX2048.0885. Indeed, prior to development of NGS methods, PCR was known to introduce issues into genetic sequence analysis due to errors and amplification bias. EX1012.0930, .0935. As explained above, Syvänen discusses the use of allele-specific primers in PCR (a method of genetic analysis using PCR). EX1012.0930. Syvänen discloses that when using such primers “without a consecutive target-specific detection reaction or size-separation step, the specificity of the method

might be hampered owing to primer dimers and other spurious amplification products that will not be distinguished from actual PCR products.” *Id.*, .0935. Syvänen notes that a “limitation of all variants of allele-specific PCR is that the reaction conditions or primer design for selective allele amplification must be optimized empirically for each SNP,” and as such, “are best suited for the analysis of a limited number of SNPs in large sample collections.” *Id.* Syvänen discloses that the technology “should avoid the PCR amplification step, and the technology could be based on the analysis of single DNA molecules, which would allow direct determination of the haplotypes formed by the SNPs.” *Id.*, .0940. In a table where Syvänen identifies the “most significant advantage,” “disadvantage,” or “special feature” of various SNP-genotyping methods, Syvänen recognizes that the most significant advantage of one technology is that “PCR amplification [was] avoided” and that a special feature of another is that it “avoids target amplification.” *Id.*, .0934.

253. Even with the advent of NGS methods, prior art publications continued to suggest that “the best way to avoid bias during PCR is to avoid library amplification by PCR altogether.” EX2042.0011. As described above, several different groups developed library preparation methods that bypassed PCR amplification. *See supra* ¶¶ 92-93. Indeed, Bronner and Quail expressly taught that “[u]nnecessary PCR amplification steps will exacerbate amplification biases, and so

they should be avoided wherever possible” and that it was possible to perform NGS without PCR amplification. EX1009.0018-19. In 2011, Quail reported that “[l]ibraries prepared without PCR amplification performed best.” EX2071.0010. Additionally, as of the early 2010s, various commercial groups had invested time and resources developing PCR-free sample preparation kits for use in NGS. *See supra* ¶¶ 92-93.

254. Given these teachings of the art, a POSA would not have been motivated to use a single target-specific primer with PCR amplification for NGS where there were not additional safeguards, such as complex adapters and microarrays. A POSA would have been concerned with issues of target specificity and off-target binding. As such, a POSA would not have been motivated to decouple the MARA method from its microarrays and to apply to NGS without additional safeguards, such as special adapters designed to prevent off-target binding.

G. Dr. Metzker’s Prior Publication Would Have Discouraged a POSA from Modifying Shapero to Practice the Claimed Inventions

255. In 2010, Dr. Metzker published an article concerning the “applications of next-generation sequencing.” EX1025.0031. In it, he “present[ed] a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments.” *Id.* He also “outline[d] the broad range of applications

for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.” *Id.*

256. Dr. Metzker included a section specific to target enrichment (what he referred to as “genome enrichment”). *Id.*, .0041. He explained that due to “whole-genome sequencing [being] still an expensive endeavor,” an “interim solution to this problem may be to use NGS platforms to target specific regions of interest.” *Id.* As Dr. Metzker explained, “[t]he concept of targeting specific regions of the genome is well established, with PCR being the most widely used method, albeit on a small scale. ***PCR coupled with high-throughput NGS platforms for targeting strategies is not practical***, as sample preparation would require handling tens of thousands of primers individually or in large multiplex groups to meet the needs of a single instrument run.” *Id.* (emphasis added).

257. The only PCR-based target capture method highlighted by Dr. Metzker as a potential solution was microdroplet PCR. *Id.* Dr. Metzker otherwise highlighted hybrid capture methods with microarrays, circularization methods (such as molecular inversion probes), and biotinylated RNA capture sequences. *Id.* Dr. Metzker concluded the section by writing that “[i]nitial reports raised concerns as to the readiness of targeted capture for routine use owing to the high degree of variability in coverage and representation, the capturing of off-target fragments, the requirement for microgram quantities of starting material and the high cost for

microarrays, but improvements are being made to address these limitations.” *Id.* None of those improvements contemplated the use of single target-specific enrichment. *See id.*

258. In view of Metzker 2010, a POSA would not have been motivated to decouple the MARA method from use with microarrays and instead use with NGS as a PCR-based method. Dr. Metzker specifically warns against coupling such PCR methods with NGS for targeting strategies given that it is “not practical.” *Id.* Dr. Metzker’s discussion of hybrid capture approaches, including recent reports of microarrays with higher than 90% capture efficiency, would be a further reason not to decouple MARA from its intended microarrays. *Id.*

259. At his deposition, Dr. Metzker referenced this section and claimed that “when [he] talk[ed] about the concept of targeting specific regions, that is traditional PCR right there.” EX2109.0017-18 (68:6-69:17). Dr. Metzker claimed that “traditional PCR” refers to “two target-specific primers amplifying a single region of interest.” *Id.* (69:8-18). Dr. Metzker does not use the term “traditional PCR” in his publication. Dr. Metzker uses the word “traditional” once to discuss “traditional fosmid-end sequencing approaches.” EX1025.0040. At no point does Dr. Metzker distinguish between other forms of PCR amplification in discussing target capture. A POSA would have understood his comment warning against using PCR with NGS target capture to apply in general to PCR-based target capture, including single

target-specific primers and primer pairs. With that said, Dr. Metzker's testimony concerning target capture reflects his view that PCR at the time was thought of as using target-specific primer pairs. This view reinforces the perspective in the art that PCR-based target capture methods used with NGS required target-specific primer pairs or more complicated safeguards. *See supra*, ¶¶ 57-62.

H. There is No Reason a POSA Would Have Decided to Modify Shapero to Practice the Claimed Methods Instead of Other Target Capture Approaches

260. As explained above, there were three categories of prior art target capture methods used with NGS: (1) PCR-based methods; (2) capture-by-hybridization; and (3) capture-by-circularization. *See supra*, ¶¶ 87-104. The claimed inventions practice an improved, novel PCR-based method. I have seen no evidence from QIAGEN or Dr. Metzker as to why a POSA would have decided to adapt Shapero to be used as a PCR-based method of NGS target capture rather than the other two methods (assuming that a POSA would even have been motivated to adapt Shapero for use with NGS, which as I have explained above, a POSA would not). Based on the prior art disclosures discussed above (especially concerning single target-specific primers and PCR amplification, *see supra*, ¶¶ 57-62), a POSA

would not have been motivated to select a PCR-based method over hybrid capture or circularization methods.

I. A POSA Would Not Have Had a Reasonable Expectation of Success in Modifying the MARA Method to Practice NGS As Practiced by the Claimed Inventions

261. As explained above, the POSA would have been aware of decades of publications concerning issues with target specificity and off-target product generation that were associated with PCR primers. *See supra*, ¶¶ 52-62. The POSA would have been aware of the various genome walking publications that pre-dated NGS and demonstrated a need for additional features to promote target specificity beyond just using simple primers (such as complicated adapters and PCR suppression). *See supra*, ¶¶ 57-62. The POSA would also have been aware of numerous publications, including from Dr. Metzker, that discussed how PCR amplification was not desirable and should be avoided with NGS. *See supra*, ¶¶ 255-59. Based on this landscape of prior art, a POSA would not have had a reasonable expectation of success in using a single target-specific primer to enrich for sequences of interest and sequencing those enriched sequences with NGS.

262. A POSA would not have expected the MARA method to be an improvement over the art that already existed given its use of a single locus-specific primer. A POSA would have instead expected that using only one-target specific primer would result in considerably lower levels of target specificity given that the

primer was more likely to bind off-target and there was no second primer to help minimize errors resulting from that off-target binding. *See supra*, ¶¶ 57-62, 90-92. Due to the higher sensitivities of NGS for off-target sequences, the POSA would have expected that considerably more of the sequencing read outputs would be noise, which would have made it even harder to genotype SNPs. *See supra*, ¶¶ 90-92, 106-10.

263. QIAGEN, however, claims that a POSA would nevertheless have had a reasonable expectation of success in decoupling the MARA method and using it instead with NGS. There is no support for that. As I have explained above and explain further below, the data reported in Shapero is inapplicable to NGS given Shapero's reliance on microarrays. Shapero provides no insight into the MARA method that would contradict the teach-aways of the art. Shapero's disclosures and the subsequent commentary of Shapero instead reinforce concerns with target specificity and off-target binding. Delseny also provides no support since Delseny does not address the use of target capture with NGS for genotyping nor the use of a single target-specific primer with NGS for any purpose.

1. Shapero's Call Rates and DS Are Inapplicable

264. QIAGEN claims that the call rates and DS reported in Shapero support the conclusion that MARA resulted in "high on-target specificity." Paper 2, 37. I disagree. As explained above, the call rates in Shapero reflect whether a SNP was

present in the sample (based on the presence of the associated loci). *See supra*, ¶¶ 162-66, 220-224. The DS reflects the binding specificity of the target-specific array probe to the specific target. *See supra*, ¶¶ 162-66, 220-224. Neither metric reveals how MARA performed in amplifying specific loci over other sequences in the sample. *See supra*, ¶¶ 216-229. Because Shapero's microarrays were designed to filter any off-target products generated by the MARA method (which is acknowledged by Shapero), there are also no inferences that can be drawn as to the population of off-target products. *See supra*, ¶¶ 160-61, 220, 226.

265. As I also explain above, Shapero's comparisons of the data for the 250- and 750-plex where locus-specific primers are used and where they are omitted cannot support QIAGEN's claims. *See supra*, ¶¶ 221-24. A POSA would have understood from Shapero's disclosed methodologies that the absence of the locus-specific primers meant that no extension or amplification occurs. *See supra*, ¶¶ 222-23. In other words, little-to-no additional sample is generated. This is important because microarrays require a minimum amount of genetic sample volume for adequate coverage and failing to achieve that threshold can result in abysmal call rates and DS. *See supra*, ¶¶ 222-25. A POSA would have understood that comparisons of adequate sample material are not the same as evaluating levels of enrichment. *See supra*, ¶¶ 222-26. There is nothing reported in Shapero that would allow for a POSA to evaluate what enrichment specifically occurred as a result of

MARA. *See supra*, ¶¶ 222-26. For these reasons, there is nothing in Shapero that would improve a POSA's expectations with decoupling MARA from microarrays and using with NGS instead.

2. Other Data Reported by Shapero Would Have Diminished a POSA's Already Low Expectations

266. Additionally, as I explain above, Shapero reported concordance rates for each set of multiplex reactions. *See supra*, ¶¶ 164-67. A POSA would have understood that these concordance rates were meant to compare using MARA with Affymetrix's microarrays to using a whole genome sampling assay with the same microarrays. *See supra*, ¶¶ 227-28. As I previously noted, the whole genome sampling assay employed by Shapero was designed specifically to avoid the use of locus-specific primers. *See supra*, ¶ 227. In reviewing the concordance rates, a POSA would have understood that using MARA with the microarrays resulted in four-to-five times higher errors than when using the whole genome sampling assay. *See supra*, ¶ 228. This result would have affirmed to a POSA the prior art view that using a single target-specific primer resulted in lower accuracy. *See supra*, ¶¶ 52-62.

267. A POSA's low expectation of success in using MARA with NGS would have been compounded by Shapero's reporting of the higher-plex reactions. When Shapero tested hybridization for reactions using more than 1,000 array probes, Shapero observed that only 75% of primers demonstrated a high SNP call rate. *See*

supra, ¶¶ 167-68. For the other 25%, Shapero could not draw a conclusion about what resulted in low SNP call rates. *See supra*, ¶ 169. The only conclusion that Shapero could reach was that “non-specific primer-primer interactions at the higher multiplex level may not solely be responsible for failed target generation.” *See supra*, ¶ 169; EX1006.0007. A POSA would have understood that Shapero could not rule out off-target binding issues. This uncertainty would have only increased a POSA’s concerns with target specificity and off-target binding for high-throughput methods like NGS.

3. Delseny’s Disclosure on Using Genotyping with NGS is Inapplicable

268. As previously discussed, Delseny disclosed that NGS “can substitute for the genotyping microarray strategy because of its higher resolution.” *See supra*, ¶¶ 186-91. A POSA would have understood that the disclosure—along with the other disclosures concerning using NGS for genotyping—was cabined to the sequencing of an entire genome ***and not target capture***. *See supra*, ¶¶ 186-91. This understanding is reflected by the only paper cited by Delseny for the proposition that NGS can be used for genotyping: Han. *See supra*, ¶ 191. In view of Han, a POSA would have understood that Delseny was not claiming that NGS could successfully be used with target capture for genotyping. Rather, a POSA would have understood that Han explicitly taught that focusing on specific loci (as Shapero did) would not work for genotyping with NGS given accuracy concerns. *See supra*, ¶¶ 192-96. If

anything, a POSA considering Delseny and Han would have had an even further diminished view of using MARA with NGS instead of a microarray.

269. A POSA would also have gained no confidence from Delseny's disclosures concerning target capture with NGS. None of these disclosures concerned the use of a single target-specific primer for enrichment. *See supra*, ¶¶ 178-85. Rather, these disclosures reinforced the need for primer pairs or complicated target capture methods, such as the padlock probe circularization method. *See supra, Id.*

270. Delseny also expressed several concerns with using NGS for SNP genotyping, including base-calling errors, read artifacts, and false discovery rates. *See supra*, ¶¶ 175, 189. These concerns led Delseny to describe pre-NGS Sanger technology as the "gold standard" given its higher read fidelity. *See supra*, ¶ 175. A POSA's concerns with target specificity and off-target binding would not be alleviated by Delseny given these disclosures and Delseny's failure to teach how to use a single target-specific primer enrichment method for NGS. *See supra*, ¶¶ 178-91.

J. How the Art Cited to Shapero Demonstrates the Limited Utility of MARA

271. In the years that followed Shapero's publication, only twenty-one articles cited to it.

272. I have included along with my declaration a list of the publications that cite to Shapero (EX2180). The list includes 27 publications as identified through a search of Google Scholar, Dialog Proquest, and the citation management program¹⁷ provided by the publishing journal Nucleic Acids Research.

273. Of those publications, four appear to be erroneously listed as citing Shapero (Wu 2005, Kim 2005, Van Baren 2006, and Hall 2007), as confirmed by manual inspection of the references. Further, two additional citations are cumulative of other cites in the list: one is a book chapter disclosing the same article as a separate citation (compare Zhou 2005 and Zhou 2016) while the other is a dissertation repeating the same disclosures as a published article by the same author (compare Bui 2009 and Bui 2009b). As such, there appear to be only twenty-one unique article citations to Shapero in the literature. Of these references, one (Vaniusheva 2005) was excluded from the below analysis as an English-language version of the document could not be acquired. Of the remaining twenty unique references, none discuss using MARA with NGS and many do not consider MARA separate from Affymetrix's microarrays. I summarize how MARA was described in the prior art below:

¹⁷ <https://badge.dimensions.ai/details/id/pub.1048265948>.

- **Dahl 2005** (EX2126): Within a year of publication of the MARA method, Dahl 2005 described MARA as one of “[a] few methods [that] exist that permit the user to amplify subsets of genomes using variants of so-called adaptor PCR (6–8).” EX2126.0001. However, Dahl 2005 specifically noted that despite the existence of the MARA method “the only way to amplify many defined sequences in the same reaction is by combining several different PCR *primer pairs*. A crucial problem with this approach is the appearance of amplification artifacts, such as primer-dimers, truncated fragments or amplification products of irrelevant target sequences.” *Id.* (emphasis added). Dahl 2005 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. Rather, Dahl proposes a method of circularized capture. *Id.* Dahl confirms the accuracy of this method by using a microarray. *Id.* The publication does not discuss NGS.
- **Dahl 2005b** (EX2177): Also in 2005, the lead author Dahl published his dissertation (“Dahl 2005b”) wherein he provided similar commentary on the MARA method. Specifically, Dahl 2005b described the MARA method as a “an approach to amplify 10-20 specific targets in parallel where each fragment is amplified using one specific primer and one general primer complementary to an adaptor attached to each target.” EX2177.0013. Dahl 2005b noted that “[t]he authors claim that the content of the amplified product is sufficiently

defined to be used for genotyping using high-density oligonucleotide arrays.”

Id. However, Dahl 2005b explained that the MARA method is limited to the array applications, stating that “even though the complex content of the amplified product suits this particular application, the amplification method still copies a large collection of undesirable DNA fragments that will generate problems in other types of DNA analysis.” *Id.* As with the prior publication, Dahl 2005b does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. Rather, Dahl proposes a method of circularized capture. *Id.*, .0002. Dahl confirms the accuracy of this method by using a microarray. *Id.*, .0026. The publication does not discuss NGS.

- **Benjamin 2005** (EX2127): Benjamin cited to Shapero as an example of where “advances in the enabling technology for higher-throughput gene association studies, such as . . . highly multiplexed genotyping on DNA microarrays, offer unprecedented ease and efficiency in the full exploitation of human genetic diversity.” EX2127.0498. Benjamin does not discuss the MARA method nor encourage its use. Rather, Benjamin praises developments in microarray technology. *Id.*
- **Zhou 2005** (EX2128): Similarly, Zhou discusses Shapero as an example of “allele-specific hybridization.” EX2128.0001. Zhou is an article entitled

“Multiplex SNP typing by bioluminometric assay coupled with terminator incorporation (BATI)” and is concerned with developing and disclosing a “novel multiplexed approach” to pyrosequencing. *Id.* Notably, despite express knowledge of Shapero, Zhou does not disclose nor even contemplate using MARA in conjunction with pyrosequencing. Zhou does not even discuss MARA. It focuses instead of SNP genotyping through allele-specific microarrays.

- **Ronaghi 2005** (EX2129) is a review article concerning SNP genotyping that directly commented on the MARA method, describing the MARA method workflow and noting that “[a]llege discrimination is achieved by hybridization to high-density DNA oligonucleotide arrays.” EX2129.0121. Ronaghi also discloses that a “[m]ultiplexing reaction of more than 4000-plex was recently reported with 75% success rate.” Ronaghi 2005 identifies only the enzymatic step of the MARA method as an improvement over whole genome sampling assays. *Id.* Ronaghi 2005 also cautions that a “higher decrease in conversion rate is expected from chip hybridization on other chromosomes” given that Shapero optimized the genotyping conditions by basing it on the well-studied chromosome 21 and choosing SNPs already in Perlegen’s database. *Id.* Ronaghi 2005 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. In fact, in discussing

DNA sequencing methods to be used for SNP genotyping, Ronaghi discloses Sanger sequencing and an upcoming “microchip-based device that can label and process DNA fragments from individual microbeads in low-volume reactions, followed by ultra-fast separation and analysis on microfabricated capillary electrophoresis channels.” *Id.*, .0112. In discussing sequencing by synthesis, Ronaghi discloses that sequencing single molecules of DNA—instead of PCR amplicons—“would eliminate the need for costly and often cumbersome procedures, such as cloning and PCR amplification, and also avoid the problem of asynchrony during extension.” *Id.*, .0113.

Ronaghi discusses “two major categories” of SNP genotyping techniques: singleplex to lowplex and highly multiplexed technologies. *Id.*, .0115. Ronaghi considers pyrosequencing, single-base extension techniques (such as mini-sequencing), mass spectrometry-based techniques, Invader assay, and Taq-man assay to be singleplex-to-lowplex. *Id.* Ronaghi explains that MIP assay and oligonucleotide ligation assay are “[m]ore advanced technologies, which allow for highly multiplex assay and enable a high level of multiplexing because they “*circumvent the amplification step prior to genotyping.*” *Id.* (emphasis added). Ronaghi explains that whole-genome sampling assay was recently developed and allows for highly multiplex SNP analysis. *Id.* In this context, Ronaghi explains that MARA is a “modification of the latter

technique” that “enables custom genotyping.” *Id.* In concluding, Ronaghi explains that “[a]n ideal technology for genotyping would identify any selected nucleotide in the genome, in all DNA samples, with high accuracy at a low cost” and that large-scale custom genotyping to find the causing haplotype will be “dominated by array-based technique.” *Id.*, .0123.

- **Howell 2006** (EX2130) described the MARA method as a variant of “so-called adaptor ligation method[s that] use[] restriction endonucleases to digest a genomic sample into fragments of a known distribution.” EX2130.0160. Howell described the variation introduced by MARA as using “one common primer[] and a single specific primer per target sequence to achieve amplification of specific genomic sequences.” *Id.* Howell discloses that such methods utilize adapters and PCR method to “[amplify] a subset of the genome [to achieve] a reduction of sequence complexity [].” *Id.* Howell does not discuss any results reported in Shapero. *Id.* Instead, Howell describes how a circularization method reported in Callow 2004 can be used for “the selective amplification of a single fragment from a genomic DNA sample.” *Id.* (citing Callow 2004, EX2176). Howell then notes that “[n]one of the methods above”—including Shapero—achieves the ideal goal of amplifying a large set of arbitrary sequences without the production of undesired amplification products.” *Id.* Howell concludes the section by highlighting

another capture by circularization method. *Id.* Howell does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method.

- **Li 2006** (EX2131): Also in 2006, Li cited Shapero for the proposition that “[h]igh-density oligonucleotide microarrays are typing systems based on hybridization assays.” EX2131.0079. Li does not discuss MARA.
- **Stenberg 2006** (EX2177): Stenberg discussed the MARA method as a “hybrid” of methods designed to “reduc[e] the complexity of the sample by directed amplification of the sequences that are targets of the analysis,” such as multiplex PCR. EX2177.0014-15. Stenberg noted that the MARA method reduced the number of primers required for target amplification, thereby alleviating issues with multiplex PCR such as primer dimers. *Id.* Stenberg further described the microarray used in Shapero as capable of distinguishing “relative number of each [allele] variant” by comparing the probe signals across probes designed to hybridize to different allele variants. EX2177.0015. However, Stenberg does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. Instead, Stenberg proposes a method of circularized capture along with a computation software tool for designing capture probes. EX2177.0002. Stenberg confirms the accuracy of this method by using a microarray. EX2177.0026-27.

- **Wang 2006** (EX2132): Wang 2006 discusses concerns with preventing “spurious amplification” resulting from multiplex PCR, and to that end designed an adapter ligation-mediated allele-specific amplification scheme that utilized y-shaped adapters designed to limit nonspecific extension products. EX2132.0240, .0242. The Wang 2006 method also employed what it termed a “suppression PCR” scheme that limited amplification from the universal adapter sequence as a method to further enhance primer specificity. EX2132.0240-41. Wang 2006 compared this approach to the MARA method, noting that in addition to the Y shape for the adapter, they could “modify the oligonucleotide at the 3’ end of the adapter to block the nonspecific extension.” EX2132.0243. Wang 2006 makes no further discussion of the MARA method. Wang analyzed the results obtained with this Y shaped adapter through size-separation gel electrophoresis and acknowledged that while the “bottleneck of the typing throughput by the method is the capability of gel electrophoresis,” higher well “microplate gel[s]” will increase throughput significantly. EX2132.0247. Wang also noted that while the throughput was “much lower” than compared to microarrays (such as Affymetrix Gene Chip System), “it [was] not necessary to use such expensive platforms for routine diagnostics in hospitals.” *Id.*

The same group used the Wang method in a second 2006 paper (“Wang 2006b”, EX2133), wherein they explained that they “used an adapter to ligate either end of the digested fragment to reduce the number of primers in multiplex PCR,” which as a result “the preferential amplification of certain targets was greatly minimized.” Wang cites to Shapero for this proposition of certain targets being greatly minimized. Wang then proceeds to discuss how the Y shape design of the adapter resulted in the suppression of “non-specific amplification in multiplex PCR.” EX2133.0223-24. Despite these comparisons, neither Wang 2006 nor Wang 2006b employ the MARA method itself, nor do they contemplate using MARA in conjunction with any NGS method.

- **Loakes 2007** (EX2134): In 2007, in a review of nucleotide species and chemical structure, Loakes described Shapero as one an application of microarrays for single nucleotide polymorphism detection, specifically single nucleotide polymorphisms. EX2134.0234. Loakes does not discuss the MARA method.
- **Morin 2007** (EX2135) described Shapero as one of several SNP analysis methods addressing issues in the field related to PCR issues and low-amount DNA samples, stating that such methods “now appear to make screening of nuclear genetic variation possible for the first time from a wide range of poor-

quality samples” and “increase throughput to allow for efficient amplification of a larger number of loci in the multiplex PCR.” EX2135.0938, .0945. Morin 2007 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. In fact, Morin 2007 cites Shapero as an example where “[a]dditional modifications will be needed to ... increase throughput to allow for efficient amplification of a larger number of loci in the multiplex PCR.” EX2135.0945.

- **Stromqvist Meuzelaar** (EX2136): Additionally in 2007, researchers continued to develop multiplex PCR strategies for SNP genotyping that sought to mitigate off-target effects endemic in high multiplex PCR methods. EX2136.0835. Stromqvist Meuzelaar noted that “[t]he key challenge in developing a high-multiplex amplification procedure is preventing excessive off-target priming by the many primers in the reactions.” *Id.* Stromqvist Meuzelaar explained that “[s]everal previously developed methods use surface immobilization of primer pairs as a way to separate reactions and thus limit primer-dimer formation, but drawbacks of these methods include inefficient amplification, loss of reactants from the surface, and considerable primer-dimer formation within pairs of primers.” *Id.* One solution to these problems reported by Stromqvist Meuzelaar was “MegaPlex PCR,” which uses surface-bound primers to mitigate primer-dimers. *Id.* Stromqvist

Meuzelaar reported that MegaPlex PCR also uses a common primer approach for universal amplification, and compared that approach to the PCR scheme used in Shapero. *Id.* Stromqvist Meuzelaar does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method.

- **Zakharova 2008** (EX2137): In 2008, in a discussion of PCR-microarray nucleotide assessment methods, Zakharova described a general approach for SNP screening using PCR amplification by one locus specific primer (per SNP) and an adapter-ligated common primer. EX2137.0002. Zakharova described Shapero as a “modification of this procedure” that “can produce specific amplification of desired fragments.” *Id.* Zakharova 2008 follows this statement up by noting that “whole-genomic DNA can be amplified without any reduction of complexity and the resultant product can be used e.g. for array-based comparative genomic hybridization.” *Id.* Zakharova explains that “[s]olid phase PCR in microarray technology provides an elegant way to solution of many micro-array related problems” given that “at least some of the immobilized oligonucleotides serve as primers in PCR carried out right over the microarray,” preventing primer dimer formation. *Id.* Zakharova does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. Instead, Zakharova performs solid-phase

PCR with a microarray using a single target-specific primer before conducting universal amplification with two liquid-phase primers. *Id.*, .0005.

- **Bui 2009** (EX2138): By 2009, MARA was considered one of “a large variety of techniques for high-throughput SNP genotyping [that] are available.” EX2138.0002. Bui described Shapero as one of “[v]arious genome re-sequencing methods [that] are also extremely powerful for large scale SNP-discrimination,” but noted that it is “impractical for assaying a selected set of SNPs in specific genomic regions, and [is] usually beyond the reach of small to medium sized laboratories with limited resources.” *Id.* Indeed, it is this issue of accessibility that motivated Bui 2009 to develop “a genotyping method named Simple Allele-discriminating PCR (SAP), which is ideally suited for the small-scale genotyping and gene mapping routinely performed in small to medium research or teaching laboratories.” EX2138.0001. As Bui explains, SAP utilizes allele-specific primers to generate PCR products only when the specific allele is present (unlike the MARA method, which generates locus-specific products that contain the allele site and then identifies the specific allele via array hybridization). EX2138.0003-04. Bui does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method.

- **Palka-Santini 2009** (EX2139) described various multiplex PCR methods used for microarray-based genotyping and noted that Shapero was an “alternative approach applied to solution-phase highly multiplex PCR [that] replace[d] target-specific primers with universal ones.” EX2139.0010. However, Palka-Santini 2009 observed that the MARA method “involve[d] multiple steps starting with enzymatic digestion of the template DNA, ligation to adapters, primer extension and finally two subsequent PCR reactions. Such multi-step approaches are time consuming and prone to contamination and therefore have not been recommended for bacteriological routine diagnostics.” *Id.* Palka-Santini then proceeded to explain that “[s]ince the presence and concentration of the particular pathogen in a microbiological laboratory is unknown, the multiplex reaction should include as many primer pairs as capture probes are present on the microarrays.” *Id.* Palka-Santini 2009 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method. In fact, Palka-Santini 2009 instead explains how “[i]n the last two years the publication of several reports referring to rapid identification of bacterial species by multiplex PCR coupled to microarrays detection demonstrated the usefulness of this approach and the growing interest in implementing it in routine diagnostics.” *Id.*, .0012.

- **Jin 2011** (EX2140) described Shapero as one of “many methods [] developed to detect single-nucleotide polymorphism.” Jin 2011 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method.
- **Liu 2012** (EX2141) classified Shapero as one of “[a] large variety of techniques for high-throughput SNP genotyping” specifically employing “genome re-sequencing.” EX2141.0002. Liu 2012 considered Shapero and similar methods to be “expensive and require specialized equipments, which cost more [than] standard primers and are not practical for assaying low- to moderate-throughput SNPs.” *Id.* As such, Liu 2012 concluded that “there is a need for simple and accurate genotyping assays that can be implemented in laboratories lacking access to sophisticated equipment.” *Id.* Despite express knowledge of both Shapero and NGS techniques, Liu 2012 does not employ the MARA method, nor does it contemplate using MARA in conjunction with any NGS method.
- **Kaur 2018** (EX2178): Kaur describes a software for the design of allele-specific primers. EX2178.0291. Shapero is listed as one of fifty-three articles and sources in the appendix of Kaur. EX2178.0295-96. Kaur does not provide in-line citations to any of the listed references, and as such it is unclear what proposition exactly Shapero is cited for.

- **Liu 2024** (EX2142): Shapero was not uniquely cited again for a further six years, until a 2024 article by Liu *et al.* EX2142.9201. In a discussion of multiplex PCR methods and associated issues, such as unequal amplification rates and primer dimers, Liu 2024 described the Shapero as a “[v]ariation[] of [a] universal primer-based multiplex PCR [that had] been successfully used for large-scale array-based single nucleotide polymorphism (SNP) discovery and genotyping assays” in considering multiplex PCR methods to further develop to achieve a “generic multiplex asymmetric PCR approach with a simple design, easy optimization, high sensitivity, and balanced amplification.” *Id.* However, Liu 2024 rejected the MARA method in favor of a homo-Tag-assisted nondimer system (HANDS) method because the HANDS method allowed for symmetric PCR – a method allowing unbiased primer function on either side of a given amplicon. *Id.*

274. Indeed, as demonstrated above, despite over twenty-one years in the literature (including ~eight before the priority date of the Tecan patents), no researcher considered applying the MARA method to an NGS platform. In fact, it does not appear as though the MARA method was ever used by another research group to generate published data. Instead, the MARA method was variously described as unable to “achieve[] the ideal goal of amplifying a large set of arbitrary sequences without the production of undesired amplification products”

(EX2130.0160), “impractical for assaying a selected set of SNPs in specific genomic regions, and [] usually beyond the reach of small to medium sized laboratories with limited resources” (EX2138.0002), “time consuming and prone to contamination” (Palka-Santini 2009.0010), and “expensive and requir[ing] specialized equipments, which cost more [than] standard primers and [] not practical for assaying low- to moderate-throughput SNPs” (EX2141.0002). As such, off-target primer binding remained an unresolved issue inherent in PCR applications. Indeed, by the late 2000s, researchers were considering alternatives to PCR amplification to circumvent these issues and enable multiplex sequencing of targeted gene regions. As such, the MARA method represented a niche, highly specialized, by-product-plagued, and expensive method that was rarely cited and rarely, if ever, employed by the academic scientific community. In my opinion, a POSA would have been unlikely to select the MARA method when considering potential methods of array-based genotyping, let alone select the method for the further extensive development required to render it compatible with an NGS platform.

X. SECONDARY CONSIDERATIONS

275. The nonobviousness of the claimed methods is reinforced by objective indicia of nonobviousness, including at least that (1) the prior art taught away from the claimed methods; (2) the art expressed skepticism towards the claimed inventions; (3) a long-felt, unmet need existed that was resolved by the claimed

inventions; (4) the industry praised the claimed inventions; (5) others copied the claimed inventions; and (6) using the claimed inventions yielded results that would have been unexpected based on the closest prior art.

A. Nexus

276. As explained above, the claimed inventions are directed to the use of a single target-specific primer (the “one or more oligonucleotides” with “at least 10 bases designed to be complementary to the nucleic acid sequence of interest”) to selectively amplify and enrich a target sequence with a pair of common adapters that can then be sequenced on a massively parallel sequencing platform. *See supra* ¶¶ 118-121. This combination of limitations is unique and characteristic for all claims of both patents. Tecan (and its predecessor, NuGEN) refer to the claimed technology as “Single-Primer Enrichment Technology” or “SPET.”

277. The specification of the patents discloses “[p]roducts based on the methods of the invention may be commercialized by the Applicants under the trade name Ovation®. Ovation® is a trademark of NuGEN Technologies, Inc.” EX1001.0031 (31:30-33); EX1009.0031 (31:30-33). Example 1 of the patents demonstrates the claimed methods through the use of NuGEN’s Ovation Ultralow Library System. EX1001.0031; EX1009.0031.

278. After the priority date, NuGEN published an advertisement in Nature’s Blood Cancer Journal where NuGEN described the technology as such:

SPET (single-primer enrichment technology) another innovative NuGEN technology, maximizes the information yield from NGS with a rapid and efficient sample-preparation workflow. SPET is a target-enrichment tool that allows for the simultaneous detection of single-nucleotide polymorphisms (SNPs), copy-numbers variants (CNVs) and insertions or deletions. Until recently, CNV and sequence analysis would have required completely different analytical platforms and sample preparation to elucidate information for these three types of genetic variation. SPET does so in a single assay, conserving patient samples and making more efficient use of sequencing resources. SPET is the core technology used in Ovation Cancer Panel 2.0 Target Enrichment System.

SPET can also be used to identify gene-fusion events and sequence variants in an RNA sample. The SPET technology for targeted RNA sequencing is flexible enough to target any gene in any organism for gene-fusion detection and gene-expression analysis. This will be an important contribution to RNA-seq-based cancer diagnostic and prognostic tests.

EX2082.0001.

279. In 2015, named inventors Jonathan Scolnick and Douglas Amorese reported on the use of SPET. EX2083.0001. In the publication, they illustrate how the Ovation Target Enrichment System utilizes SPET.

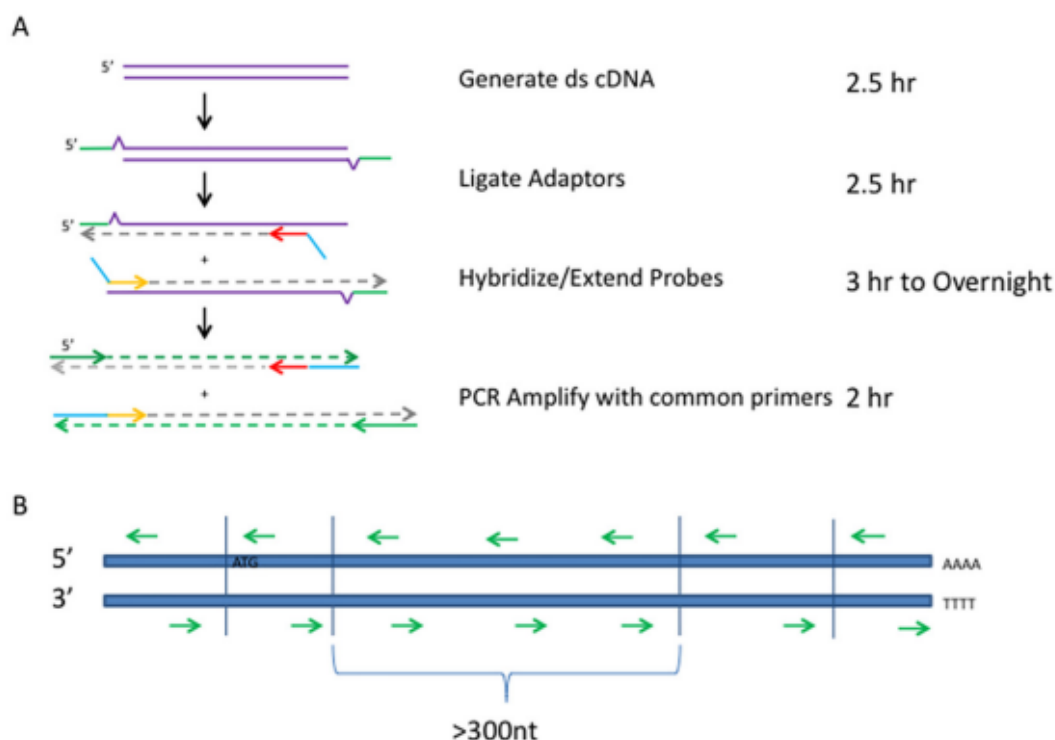


Fig 1. Description of the Ovation Target Enrichment System. (A) Experimental steps of the assay and time required for each step. Adaptors (green) are ligated on to generated double stranded cDNA (ds-cDNA). Probes (shown in red and yellow) are hybridized to target cDNA and extended with a polymerase (dashed grey lines). All probes have common tail sequences (blue), which are used as priming sites along with adaptor sequences in subsequent library amplification PCR steps. (B) Example of probe positioning across different exons in a full length double stranded cDNA. Each exon (demarked by blue vertical lines) will have probes (green arrows; arrow points in the 3' direction) designed to hybridize near the predicted exon-exon junctions. Exons larger than 300 nucleotides (nt) may have additional probes tiled along the length of the exon to obtain more complete sequence coverage. Probes are designed against both strands of the cDNA to enable identification of gene fusions when only one of the pair of genes is targeted. Translation start sequence (ATG) and poly A tail are labeled.

doi:10.1371/journal.pone.0128916.g001

Id., .0002.

280. This illustration is coextensive with steps 1(a) through 1(f) of the '012 patent and 1(a) through 1(c) of the '108 patent. The publication explains that the “resulting library was sequenced on an Illumina MiSeq,” satisfying the last steps of both independent claims.

281. In 2017, NuGEN announced the issuance of U.S. Patent No. 9,650,628 (“the ’628 patent”). The ’628, ’012, and ’108 patents all claim priority to the same provisional application. *See supra* ¶ 136. NuGEN explained at the time that the ’628 patent describes “the Single Primer Target Enrichment Technology (SPET) which enables targeted nucleic acid analysis by next generation sequencing (NGS). NuGEN offers commercial solutions using SPET to create target and strand-specific libraries ready for NGS.” EX2084.0001. NuGEN stated at the time that “[p]roducts that SPET is currently incorporated [include] Ovation® Custom Target Enrichment System, Ovation® Custom Gene Fusion Target Enrichment System, and Allegro Targeted Genotyping by Sequencing (TGBS).” *Id.*

282. After Tecan acquired NuGEN, Tecan incorporated SPET into the Allegro® Targeted Genotyping V2. According to its product sheet, it “us[es] the patented Single Primer Enrichment Technology (SPET) approach to specifically target SNPs of interest.” EX2085.0001. It notes that this technology provides for “[h]ighly scalable multiplexing to enable largescale genotyping.” *Id.*

283. The Allegro Targeted Genotyping V2 product sheet also depicts its workflow:

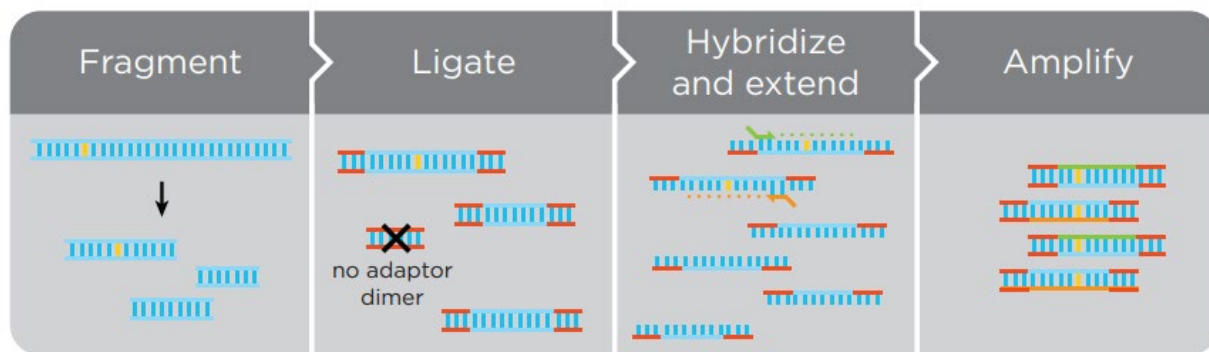


Figure 1: Allegro Targeted Genotyping V2 is a complete end-to-end solution for targeted genotyping, with a simple easy to follow workflow.

Id.

The product workflow demonstrates that the product is coextensive with the claimed inventions. Nucleic acids are fragmented, and the nucleic acid fragments are ligated to partial duplex adapters. An oligonucleotide complementary to a nucleic acid sequence of interest is annealed to the sequence of interest and extended. These extension products are amplified using primers directed to the adaptor sequences to generate libraries of amplified products containing the oligonucleotide extension product sequence or its complement. As explained on the product sheet, this library preparation method generates amplicons that can then be sequenced using NGS.

Id., .0001-02.

284. Based on the patents, the prior statements made by Tecan and NuGEN, and the product manuals illustrating the commercial embodiment of SPET, it is my conclusion that the Tecan Ovation and Allegro products embody the claims of the '012 and '108 patents.

B. Teach-Away/Skepticism

285. The methods claimed in the '012 and '108 patents disclose the use of a single target-specific primer to generate oligonucleotide extension products containing a sequence of interest and the amplification of the oligonucleotide extension products to generate enriched sequences of interest that can be sequenced through NGS. In my opinion, as of 2012, a POSA would have been aware of the nearly two decades worth of literature that demonstrated concerns with amplification, such as target specificity and off-target binding. *See supra* ¶¶ 57-62. Additionally, a POSA would have been aware from literature concerning the development and improvement of NGS library preparation methods that the need for accurate target specificity was paramount. *See supra* ¶¶ 87-104.

286. For example, numerous pre-NGS genome walking papers taught that there were issues associated with using a single target-specific primer with PCR amplification given the generation of off-target products that would confound subsequent sample analysis. *See supra*, ¶¶ 57-62. The art over time reflected the need for either primer pairs or complicated safeguards when using PCR-based target capture methods. *See id.* Numerous references discouraged the use of PCR amplification methods, such as MARA, for NGS and genotyping, including Syvanen and Metzker. *See supra*, ¶¶ 63-64, 172-73, 236, 255-59. The commentary concerning Shapero disclosed MARA's issues with off-target product generation and

emphasized the further use and development of microarray technologies. *See supra*, ¶¶ 271-74. Not one reference in the nearly eight years after Shapero published and before the priority date suggested that MARA should be decoupled from its microarrays or used with NGS. *See id.*

287. A POSA would have understood that the field was skeptical of the specificity of single target-specific primer methods due to the generation of off-target products. *See supra* ¶¶ 57-62. A POSA would be particularly skeptical of coupling a single target-specific primer method with an NGS platform, as NGS platforms already suffered from issues related to PCR bias. *See supra* ¶¶ 63-64. Indeed, a POSA would have understood from the art at the time that most single target-specific primer methods employed complicated experimental parameters designed to reduce off-target primer binding, including complicated vector schemes, chemical modification strategies, non-target degradation treatments, and capture instrumentation. *See supra* ¶¶ 57-62. Moreover, a POSA would have appreciated that most of those solutions served only to reduce off-target effects rather than eliminate them. *Id.* As such, the prior art would have taught away from the claimed inventions and cast skepticism on various parts of the claimed inventions.

C. Long-felt, unmet need

288. Despite the benefits conferred by the massively parallel nature of NGS, researchers quickly recognized that early NGS methods suffered from “the inability

to easily target specific regions of a genome.” EX2062.0003. As such, researchers recognized that new “sequence capture methodologies [were] required for time- and cost-effective sequence analysis in high-throughput genomic projects.” EX2064.0001. Such improvements would need to be accurate, efficient, and modular. Indeed, researchers in the field observed that while the “ease of any given enrichment method and its potential for near-term implementation in a production setting are also considerations,” “the quality of the NGS data produced with a given enrichment method on a given sample type will ultimately dictate the strategies we elect to further optimize.” EX2063.0448. Further, early PCR-based target enrichment methods used in conjunction with NGS methods were known to suffer from accuracy issues. *See supra* ¶¶ 63-64, 90-99. For example, Kelley reported that even with primer pairs, they experienced unintentional amplification of unwanted sequences, which led to them discarding half of their sequencing reads. EX2179.0545 (explaining that the primer pairs used did not have the desired specificity to ensure that only the target sequences were being amplified).

289. Other early NGS-compatible target capture methods relied on specialized instrumentation, such as RNA library “bates” (short RNA molecules designed to selectively pull nucleic acid fragments of interest out of a sample) and microfluidic microarray chips (tiny silicon or glass hybridization arrays), to use, which made the methods technical and complex. EX2062.0004; EX2064.0001-02.

Indeed, in some instances, the lack of methods not requiring specialized instrumentation resulted in researchers re-using instruments, including hybridization arrays, despite the risk of sample contamination. EX2051.0053. Researchers recognized at the time that “there is greater demand for cost-effective, streamlined library preparation methods.” EX2056.0003. One such commercial solution was developed by Nextera, which sought to improve library preparation efficiency by streamlining sample preparation steps and limiting PCR amplification. *Id.* However, the Nextera method utilized a nonspecific transposase method that indiscriminately fragmented and tagged nucleic acid sample fragments. *Id.*, .0003-04. As such, the Nextera method was not viable when experimental parameters required a target capture approach.

290. In addition to the lack of accurate, efficient, and user friendly methods, PCR-based target enrichment methods suffered from logistical issues that complicated applying such methods at scale. EX2010.0002. Such methods were viewed as “laborious, time-consuming and costly,” particularly when employed in a multiplexed fashion, where it was “difficult to check successful amplification per fragment.” EX2051.0052. Moreover, such methods increased the “chance of obtaining artifacts.” *Id.* Alternative capture methods, such as capture-by-circularization, suffered from issues with consistent target capture and therefore produced variable sequencing results. *See* EX1001.0016 (2:18-24). Such methods,

similar to traditional multiplexed PCR, where each target required two target specific primers or probes, were ill-suited for use with massively parallel NGS methods which could sequence far more nucleotide sequences than could practically be targeted. EX2010.0002.

291. As of late 2011, targeted enrichment, including PCR-based methods, capture-by-circularization, and hybrid capture, all suffered from issues with “uniformity, efficiency of coverage and off-target capture.” EX2169.0007; *see also* EX2050.0001-03. Indeed, some researchers concluded that “it may ultimately be more cost-effective to sequence a whole genome” rather than use target capture. EX2169.0007. Researchers noted that target enrichment and sequencing library preparation methods were “laborious, time-consuming, prone to experimental errors and difficult to automate” and had complex protocols which require large amount of starting genomic DNA.” EX2052.0002. Such methods were further “susceptible to coverage biases secondary to amplification and molecular artifacts.” *Id.*

292. These long-felt, unmet needs are the same reported by the patents: “[t]here is a need for improved methods for selective target enrichment that allow for low-cost, high throughput capture of genomic regions of interest without specialized instrumentation.” EX1001.0016; EX1009.0016 “Additionally, there is also a need for high efficiency nucleic acid library generation.” *Id.* As the patents explain, the “methods of the invention described herein fulfills these needs.” *Id.*

D. Industry Praise for Tecan's Solution to these Long-Felt, Unmet Needs

293. Over the years, there have been numerous articles published in the industry praising the claimed inventions and how the claimed inventions addressed the above long-felt, unmet needs.

294. Barchi *et al.* is a research paper published in 2019 by researchers at four institutions in Italy and one in Spain. In it, the authors test NuGEN's SPET technology. Specifically, the authors identify this technology by stating that "[r]ecently Nugen® developed the single primer enrichment technology (SPET, United States Patent 9,650,628)." EX2065.0002. They report using NuGEN's Allegro® SPET products. *Id.* In the publication, they describe SPET as "a new, robust, and customizable solution for targeted genotyping" and "a customizable solution for targeted sequencing at an affordable price." *Id.*, .0001-02. They explain that the technology's "reliance on single primers, rather than on primer pairs, greatly simplifies panel design, and allows higher levels of multiplexing than PCR-based genotyping." *Id.*, .0001. They conclude by stating that the technology "perform[ed] well with DNA samples prepared by different laboratories using different DNA mini-preparation protocols," that it provided "the ability of multiplexing thousands of samples in a single sequencing run," and that it "allow[ed] the discovery of thousands of novel SNPs not originally included in the panel." *Id.*, .0012-13.

295. Caitlin Smith, who obtained her Ph.D. in neuroscience Yale University and frequently authors publications concerning genomics and neuroscience research in Biocompare's Future Lab, published an editorial article in 2021 discussing the claimed technology. Smith reports on "target enrichment using single primers, supported by Tecan Genomics' Allegro® Target Genotyping V2 with Single Primer Enrichment Technology (SPET)." EX2067.0003. Smith describes the technology in this way:

Enzymatically fragmented sample DNA is ligated to adaptors containing barcodes with their DimerFree® technology. The barcoded samples are pooled and then hybridized to probes/primers, which are extended and the results amplified with PCR to create a targeted library.

Id. Smith explains that "[u]sing single primers instead of primer pairs greatly simplifies the design of primer panels, and allows higher levels of multiplexing compared to amplicon-based enrichment." *Id.*

296. Scaglione *et al.* is a research paper published in 2019. It reports on a study using SPET technology that was conducted by researchers at four Italian institutions and one British institution. In it, the researchers "evaluate single-primer enrichment technology (SPET) reproducibility." EX2068.0545. They report on using the Allegro targeted genotyping kit and Ovation Target Enrichment kit, which they explain are "based on the same SPET technology." *Id.* They explain that "[u]nlike other methods, [the claimed] technique allows for the execution of cost-

effective genotyping experiments while providing full control on target sites,” which it attributed in part to the “high efficiency enrichment system” and “the scalability to tens of thousands of probes in a single reaction.” *Id.*, .0548. They report that the study “results suggest that [the claimed] SPET genotyping is a very effective approach for genotyping a large number of loci,” and “suggest[ed] that it might be a valuable tool for large-scale genotyping studies in virtually all species.” *Id.*, .0550.

297. Benjamino *et al.* is a research paper published in 2021. The authors all belong to the Jackson Laboratory for Genomic Medicine, which is a prominent and well-respected research laboratory. In the paper, they reported using the Allegro Targeted Genotyping kit. EX2069.0002. While they report adapting the “single primer extension reaction that is widely used for genotyping [] and implement[ing] it as a quantitative, straightforward, and cost-effective method for profiling mouse microbial communities,” their method still employs the same patented SPET technology. *See id.*, .0002-03, Fig. 1. They ultimately described using the SPET technology in this way “as a quantitative, straightforward, and cost-effective method for profiling mouse microbial communities.” *Id.*, .0002.

298. A POSA would have understood from the industry praise that the claimed inventions succeeded in providing an effective method for using single target-specific primers with universal adapters to selectively amplify and enrich sequences of interest that could then be sequenced through NGS, despite the

concerns in the art with using a single target-specific primer and PCR amplification either together or even separately.

E. QIAGEN’S SPE Products Are Copies of Tecan’s SPET Products

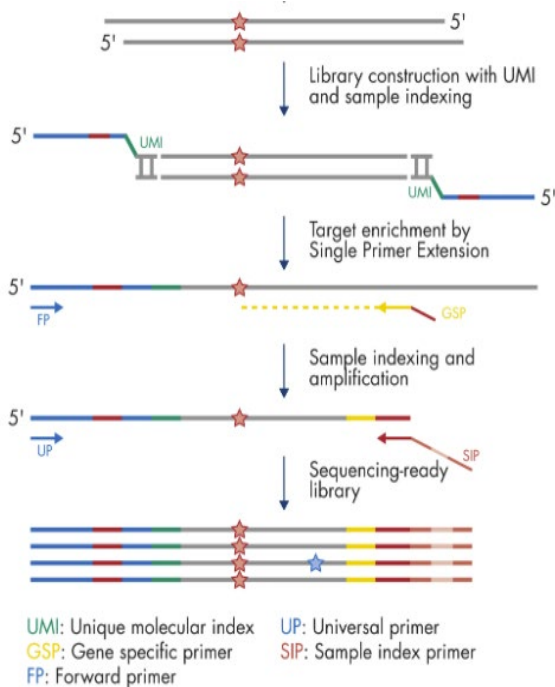
299. QIAGEN has produced a number of products in recent years under its Single Primer Extension (“SPE”) line. I understand that many of these products have been accused by Tecan as infringing the claims of the ’012 and ’108 patents. I have been asked by Tecan to consider whether some of these products are copies of NuGEN/Tecan’s commercial embodiments based on publicly available information. As explained further below, it is my opinion that QIAGEN’s SPE products, including QIAGEN’s QIAseq Targeted DNA panels and Targeted RNAscan panels, use a near-identical, if not identical, workflow to that of the claimed SPET workflow that was embodied by NuGEN’s Ovation Target Enrichment system. In so doing, I conclude that QIAGEN’s SPE workflow is a copy of Tecan’s SPET workflow.

300. As explained above, NuGEN’s Ovation Target Enrichment system is a commercial embodiment of the claimed inventions. *See supra*, ¶¶ 276-84. NuGEN published its claimed SPET workflow for its Ovation Target Enrichment system no later than July 1, 2015, after having submitted the article for publication on January 5, 2015. EX2083.0001. NuGEN published this article on its Ovation Target Enrichment system in PLoS ONE, which is regarded as a reputable, peer-reviewed open source journal. The article published approximately two years after the

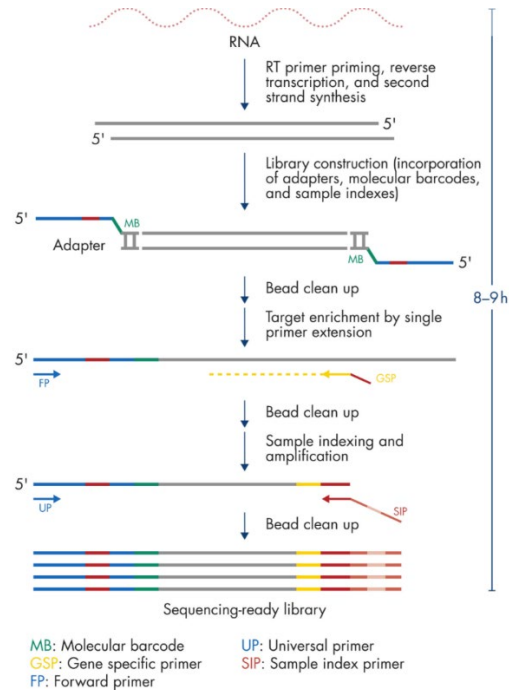
application for the '628 patent published. EX2171. It is my opinion that given the prior patent application publication and the PLoS ONE publication, the industry would have been aware of NuGEN's SPET workflow by 2015.

301. From my review of the publicly available materials, QIAGEN released its SPE products after this publication date, including its QIAseq Targeted DNA and RNAscan Panels. I have reviewed the SPE workflows associated with both products (reproduced below) and have concluded that the workflows are near-identical (if not identical) except for the sample inputs (the sample input is DNA for one product and RNA for the other).

QIAseq Targeted DNA Panel



QIAseq Targeted RNAscan Panel

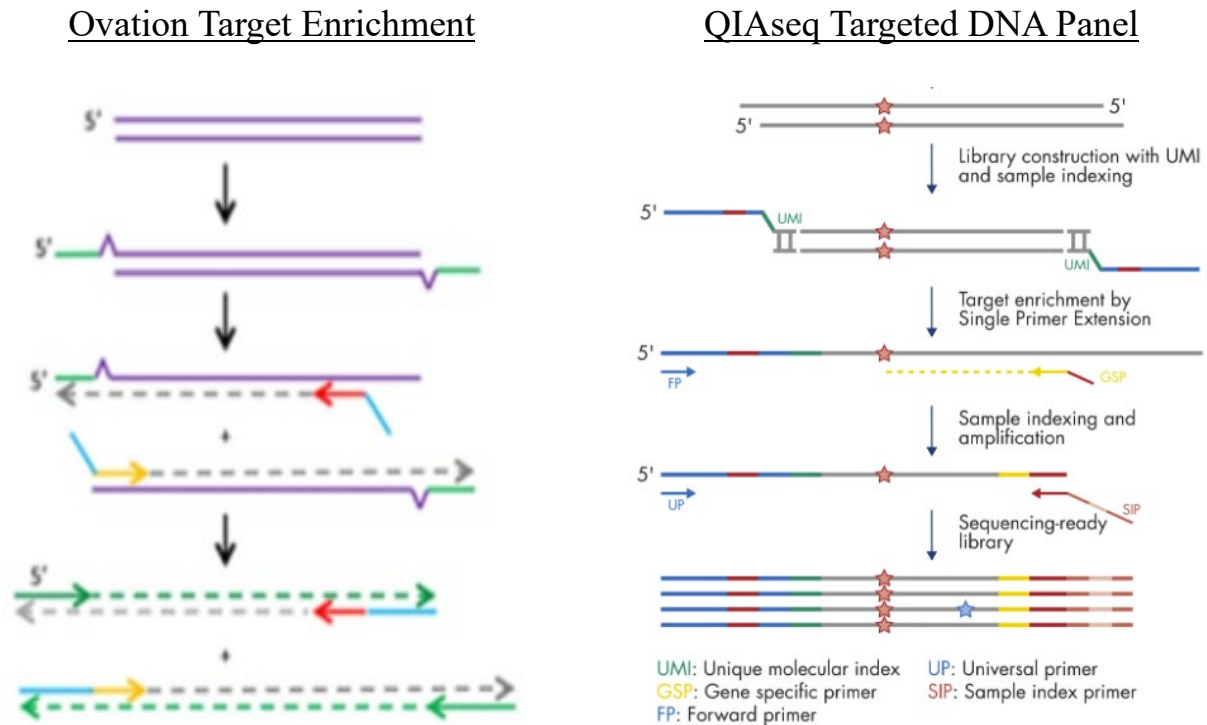


EX2143

EX2144

From the depictions, both workflows ligate partial duplex adapters to nucleic acid fragments. Both workflows also use a single-target specific primer (“GSP”) to target specific nucleic acid sequences. Both workflows depict the extension of the target specific primers and subsequent amplification using a universal primer pair (“UP” and “SIP”). And both workflows describe the resulting amplicons as a “sequencing-ready library.” It is my opinion that the SPE workflow depicted for the QIAseq Targeted DNA panels and Targeted RNAscan panels are near-identical, if not identical, and that the workflows contain the same key features for NGS library preparation.

302. With that understanding, I have also compared the SPET workflow published by NuGEN in July 2015 to the SPE workflow depicted for the QIAseq Targeted DNA panels. As reflected in the below depictions, it is my opinion that the workflows are near-identical, if not identical.



EX2083.0002

EX2143.####

Both workflows begin with nucleic acid fragments. *Id.* Partial duplex adapters are ligated to the ends of the fragments. *Id.* These adapters contain the first adapter sequence. *Id.* Oligonucleotides with target-specific primer sequences are annealed to the target sequence and extended to generate oligonucleotide extension products. *Id.* The oligonucleotide extension products are amplified using two primers specific

to the first and second adapter sequences. *Id.* The amplified libraries are prepared for NGS. *Id.*

303. Based on my comparison of these workflows, it is my opinion that the SPE workflow depicted for the QIAseq Targeted DNA panels is near-identical, if not identical, to the SPET workflow depicted for NuGEN's Ovation Target Enrichment system. It is also my opinion that the SPE workflow for the QIAseq Targeted RNAscan panels and any other products with the same or substantially similar SPE workflow would be near-identical, if not identical, to NuGEN's SPET workflow.

304. I have seen no records to suggest that QIAGEN's SPE workflow was published before NuGEN/Tecan's SPET workflow. In fact, from the records I have seen, NuGEN published its claimed SPET workflow years before QIAGEN released its SPE workflow, including in the patent application publication for the later-issued '628 patent and in its July 2015 publication concerning Ovation. As such, it is my opinion that NuGEN developed this technology before QIAGEN, that QIAGEN would have had access to these public documents from NuGEN, and that the SPE workflow of the later-released QIAGEN products is a near-identical, if not identical, copy of NuGEN's SPET workflow, which as I explained above, is an embodiment of the claimed inventions.

F. QIAGEN's Prior Statements Concerning its SPE Products

305. Because I have concluded that QIAGEN products employing its SPE workflow copy the claimed SPET workflow, it is my opinion that QIAGEN's statements concerning the SPE workflow are also applicable to the claimed inventions. I have reviewed various publications by QIAGEN, including on its public website, and have included its relevant statements below along with an explanation of their significance.

306. August 2016 – Exploring new frontiers with next-generation sequencing (EX2146)¹⁸:

- a. Describing “single primer extension” as “new” for QIAseq Targeted RNAscan Panels and explaining that it benefits users because it allows for the detection of “both known and novel

¹⁸ I note that I was able to view this webpage through a web capture from July 10, 2020, that was archived by Wayback Machine at <https://web.archive.org/web/20200710180449/https://www.qiagen.com/nl/resources/download.aspx?id=4389e6ff-3856-414e-92fd-834ff3e67715&lang=en>. A POSA would be familiar with Wayback Machine and understand that it is a reputable source for archived webpages.

fusion genes.” *This statement reflects an improvement of the claimed inventions over the prior art.*

- b. Describing “single primer extension” as “new” for QIAseq Targeted DNA panels and explaining that it benefits users because it provides “enhanced uniformity of enrichment and sequencing.” *This statement reflects improvements of the claimed inventions over the prior art.*

307. January 29, 2018 – QIAseq and Single Primer Extension Chemistry (EX2147)¹⁹:

- a. “Multiplex PCR and nested PCR-based library preparation approaches present primer design and specificity challenges inherent to these methods.” *This statement aligns with my explanation of the long-felt, unmet needs in the art.*

¹⁹ I note that I was able to view this webpage through a web capture from September 1, 2021, that was archived by Wayback Machine at <https://web.archive.org/web/20210901101209/https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/science-matters/genomics/qiaseq-and-single-primer-extension-chemistry>. A POSA would be familiar with Wayback Machine and understand that it is a reputable source for archived webpages.

- b. “Amplicon-based next-generation sequence (NGS) assays offer many advantages for targeted enrichment. However, typical approaches have suffered due to restraints in multiplexing capacity and introduction of additional bias. As primer multiplexing and PCR cycle numbers are increased, so does the potential for bias and error.” *This statement reflects the concerns in the art with amplification bias and primer dimers.*
- c. “To overcome these challenges, QIAseq NGS panels ... use single primer extension (SPE) technology which provides design flexibility and highly-specific target enrichment.” *This statement acknowledges how the claimed methods resolved these long-felt, unmet needs.*
- d. “PCR enrichment efficiency using one primer is also better than conventional two primer approach, due to the absence of an efficiency constraint from a second primer.” *This statement reflects how the claimed inventions were an improvement over the “conventional” approach.*
- e. “Single primer extension also permits discovery of unknown structural variants, such as gene fusions.” *This statement reflects another improvement of the claimed inventions over the prior*

art, including Shapero, which could not be used for unknown structural variants given its reliance on known loci and microarrays.

308. February 2018 – QIAseq SPE Technology for Illumina: Redefining amplicon sequencing (EX2148)²⁰:

- a. “[C]urrent amplicon-based enrichment and sequencing methods that rely on either two-primer designs or nested PCR approaches. ...” *This statement reflects QIAGEN’s own belief as of 2018 that the prior amplicon-based enrichment methods required primer pairs.*
- b. “two-primer designs or nested PCR approaches suffer from the following inefficiencies:

²⁰ I note that I was able to view this webpage through a web capture from June 21, 2020, that was archived by Wayback Machine at <https://web.archive.org/web/20200621214659/https://www.qiagen.com/us/resources/download.aspx?id=b3363886-aaed-4e0d-8d4b-3291b28593c5&lang=en>. A POSA would be familiar with Wayback Machine and understand that it is a reputable source for archived webpages.

- i. Low uniformity and lack of coverage due to primer dropouts
 - ii. Requirement for multiple primer pools resulting in increased DNA input per sample (see Figure 1)
 - iii. Need of a large contiguous stretch of the targeted regions to cover two primers for a nested PCR design (see Figure 2)
 - iv. Inflexibility with regards to adding primers to increase panel content
 - v. Low library complexity as both start and stop positions are defined by primers (see Figure 3)
 - vi. Inefficiency to cover entire CDS regions.” *This statement reflects numerous long-felt, unmet needs in the art, including the ones I previously identified.*
- c. **“Single primer extension (SPE) is a revolutionary approach that redefines amplicon-based enrichment and sequencing.”**
This statement reflects the nonobvious nature of the claimed inventions.
- d. “SPE overcomes the challenges of traditional amplicon-based approaches of two-primer and nested PCR designs. ...” *This*

statement confirms that the claimed inventions overcame the long-felt, unmet needs in the art. The statement also reflects QIAGEN's views as to what traditional amplicon-based approaches were (not MARA).

- e. “SPE overcomes the challenges of traditional amplicon-based approaches ... to deliver the following advantages:
 - i. Uniform coverage by reducing the number of required primers and pools
 - ii. Comprehensive coverage due to flexibility in primer design
 - iii. Ability to easily increase panel content
 - iv. Generation of complex libraries
 - v. Efficient coverage of both hotspots and entire CDS region
 - vi. Customization made easy.” *This statement reflects numerous advantages of the claimed inventions over traditional amplicon-based approaches.*
- f. “SPE utilizes a reduced number of required primers resulting in highly uniform enrichment and sequencing due to a reduction in primer dropouts (Figure 5).” *As explained further below, this*

statement reflects an unexpected result in using a single target-specific primer.

- g. “As opposed to traditional two-primer amplicon and nested PCR approaches, SPE provides a high degree of primer design flexibility, allowing for primers to be aggressively tiled across targeted regions.” *This statement reflects another improvement over prior art amplification methods.*
- h. “Because SPE requires only one region-specific primer, panel content can be easily increased to accommodate new scientific and biological findings (Figure 8).” *This statement also reflects an improvement over prior art amplification methods.*
- i. “Since each targeted region is defined by one fixed, region-specific primer, and one universal primer, the resulting library is more complex than libraries generated with traditional two-primer amplicon approaches (Figure 7).” *This statement also reflects an improvement over prior art amplification methods.*

309. The above statements by QIAGEN concern the use of a single target-specific primer to enrich for sequences of interest that can then be sequenced with NGS. It is my opinion that these statements support the nonobviousness of the claimed inventions at least because they reflect that the prior art methods of using

primer pairs extended beyond the priority date for the '012 patent, that there existed numerous long-felt, unmet needs in the art that remained unaddressed until the claimed methods, and that the claimed methods resolved these issues. The statements further reinforce that QIAGEN's SPE workflow is a copy of the claimed SPET methods and reflect further industry praise for the "revolutionary" nature of the claimed methods.

G. QIAGEN Previously Demonstrated Unexpected Results Using the Claimed Inventions

310. I understand that one hallmark of nonobviousness is that the claimed inventions yielded unexpected results as compared to what a POSA would consider to be the closest prior art. I understand that for this proceeding, QIAGEN claims that Shapero would be the closest prior art. I disagree. As explained above, Shapero reflects the use of a microarray. *See supra*, ¶¶ 162-96. The target specificity conferred by microarrays eliminates many of the off-target binding concerns that were associated with NGS and PCR amplification. *See supra*, ¶¶ 111-12. Shapero even acknowledges that the microarrays used had the capacity to handle any off-target product generation and that the target-specific probes of the microarrays enabled allele-specific discrimination. *See supra*, ¶ 161. Given the unique nature of the microarray, a POSA would not have viewed Shapero as the closest prior art.

311. Instead, a POSA would have looked to a method for sequencing an enriched nucleic acid sequence of interest with NGS, which is not Shapero. To the

POSA, it would have made sense to instead consider NGS library preparation methods given that those methods could be compared to the claimed methods to determine whether they addressed the concerns in the art, especially concerning target specificity and off-target binding. In so doing, a POSA would have sought to compare as close as possible of a prior art method. For example, a POSA would look for an amplification-based enrichment method that relied on target-specific primers (as opposed to hybrid capture or circularization). In so doing, a POSA would have concluded that the closest prior art would be target capture methods that employ the use of target-specific primer pairs to enrich for sequences of interest that are then sequenced with NGS.

312. For the reasons I have explained above, a POSA would have believed that the use of a single target-specific primer over primer pairs (or more complicated methods) would have resulted in lower coverage due to off-target binding issues. *See supra*, ¶¶ 57-62, 285-87. For example, as Miura explains, a POSA would have expected target-specific primer pairs to confer greater accuracy given the harmonious use of the primer pair to target and amplify sequences of interest. *See supra*, ¶ 61.

313. I am aware of at least one study that compared the use of a single target-specific primer and target-specific primer pairs. That study, conducted by QIAGEN,

published unexpected results demonstrating the superior performance of single target-specific primers over target-specific primer pairs.

314. On June 7, 2018, QIAGEN published an article entitled “Having issues with primer dropouts in your 2-primer amplicon-based panel? You should!” EX2149.²¹ In that publication, QIAGEN discusses what it described as “not an uncommon problem for panels that utilize a 2-primer amplicon approach for enrichment”: coverage gaps due to primer dropouts. *Id.* A POSA would understand a “primer dropout” to refer to the phenomenon where for various reasons, the PCR primers do not function as intended and do not generate amplicons from the respective target region. EX2168.0295, .0297. QIAGEN describes one such situation: “the interaction between the forward and reverse primers targeting a specific region to be enriched. If these primers share common sequences, there is a chance they bind to each other, thereby failing to enrich the targeted region, and resulting in coverage gaps (see figure).” *Id.* Other reasons can include when primers bind to off-target sequences, when primers unintentionally folding into themselves (perhaps forming hairpin structures), and when mismatches in the primer and

²¹ A version of this article appears on QIAGEN’s current webpage at <https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/science-matters/genomics/having-issues-with-primer-dropouts>.

binding sequences occur. *See* EX2048.0885; EX2047.4363, .4366-68; EX2168.0295, .0297.

315. In its June 2018 publication, QIAGEN advocates for the use of “only one region-specific primer that targets one end of the targeted region. The other primer is a universal primer that can be shared by all the region-specific primers by binding to universal sequences introduced through library adapters, an approach known as Single Primer Extension (SPE).” EX2149. While this method may have been expected to reduce instances of primer dimers, a POSA would have not expected the reduction in target specific primers to lead to better results with primer dropout given concerns with off-target binding and other potential binding issues. *See supra*, ¶¶ 57-62, 285-87.

316. Unexpectedly, QIAGEN reported that “SPE-based amplicon designs alleviate the restrictions of 2-primer-based amplicon designs, resulting in enhanced coverage (see table below), and provide extraordinary flexibility, allowing additional targets to be included without the need for a panel re-design.” EX2149. The table referred to in QIAGEN’s publication is reproduced below:

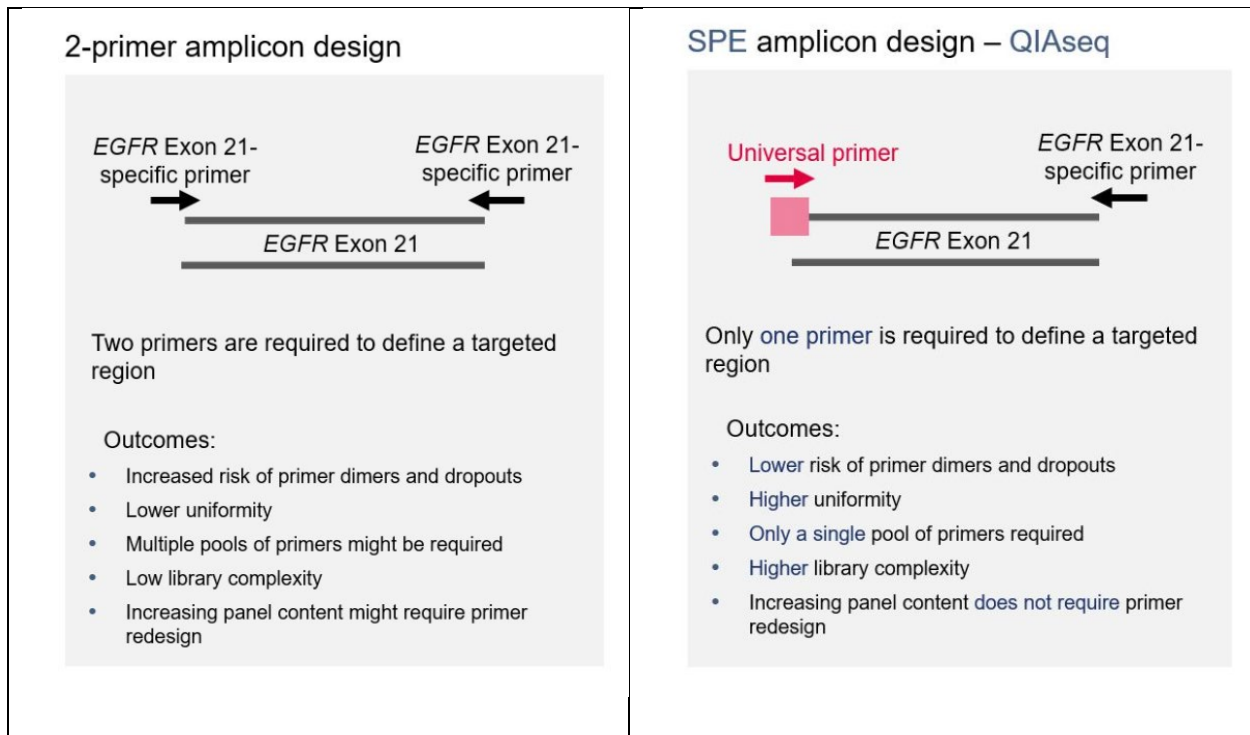
Gene	2-primer amplicon		SPE in QIAseq		
	Region	Coverage	Region	Coverage	LOD
<i>Akt1</i>	Chr14: 105246446–105246583	0	Chr14: 105246245–105246553 NM_001014431	5734	0.4%
<i>FGFR3</i>	Chr4: 1806082–1806187	0	Chr4: 1806057–1806247 NM_000142	4406	0.5%
<i>NOTCH1</i>	Chr9: 139399338–139399447	0	Chr9: 139399125–139399556 NM_017617	5406	0.4%
<i>STK11</i>	Chr19: 1220481–1220603	0	Chr19: 1220372–1220504 NM_000455	5286	0.4%
<i>TP53</i>	Chr17: 7579351–7579485	0	Chr17: 7579312–7579590 NM_001276761	5740	0.4%
	Primer dropout		No primer dropout		

As can be seen from the table, there are five genes where QIAGEN observed zero coverage when using two target-specific primers with NGS. *Id.* Conversely, for single target-specific primers, QIAGEN reported that the primers could cover a wide range of variants (thousands). *Id.* QIAGEN also reported low limits of detection (“LOD”), meaning that the assays used with the single target-specific primers were capable of detecting minute traces of the target. *Id.* A POSA would understand that primer design contributes to LOD values, with worse primer design leading to higher LOD values.

317. A POSA would not have expected, especially by the priority date, that a single target-specific primer would perform better than target-specific primer pairs. It would have also been unexpected that the use of a single target-specific primer

would result in no primer dropouts where target-specific primer pairs resulted in primer dropouts.

318. Based on these unexpected results, QIAGEN reached the following conclusions concerning target-specific primer pairs and single target-specific primers:



Id., .0003-04. These conclusions reflect the advantages of the claimed methods over the prior art.

319. It is my opinion that the unexpected results obtained by QIAGEN when it used its copied SPE methodology with NGS compared to when it used two target-specific primers with NGS evidences the nonobviousness of the claimed inventions. It is also my opinion that QIAGEN's decision to compare SPE to target-specific

primer pairs with NGS rather than microarray methods, such as Shapero, evidence its understanding that the closest prior art is target-specific primer pairs used for NGS library preparation.

XI. CONCLUSION

320. In summation, Shapero and Delseny fail to disclose each element of the independent claims of the '012 and '108 patents. A POSA would not have been motivated to combine the two references to practice the claimed inventions. A POSA also would not have had a reasonable expectation of success in doing so. Because Shapero in view of Delseny fails to render obvious the independent claims of the '012 and '108 patents, Shapero and Delseny necessarily fail to do the same for the dependent claims of the patents.

321. Tecan's inventions are novel and non-obvious. The named inventors devised the claimed methods despite the weight of the prior teaching away from and expressing skepticism towards the path they took. By inventing the claimed methods, the named inventors addressed several long-felt, unmet needs, which is reflected in the industry praise of the claimed inventions. QIAGEN's own copying and praise of the claimed inventions reflect their novelty and non-obviousness.

322. I declare under penalty of perjury that the foregoing is true and correct.

Executed on August 20, 2025, at Oklahoma City, Oklahoma.

A handwritten signature in black ink, appearing to read 'Carlos D. Bustamante', written over a horizontal line.

Carlos D. Bustamante, Ph.D.

EXHIBIT A

Curriculum Vitae

Biographical Information

Name: Prof. Carlos D. Bustamante, Ph.D.

Current positions: Founder, President and Chief Executive Officer, Galatea Bio, Inc.

Adjunct Professor of Biomedical Data Science, Genetics, and (by courtesy)
Biology, Stanford University

Venture Partner, F/Prime Capital Partners

Updated: March 1, 2024

Business Contact information

Galatea Bio, Inc.
14350 Commerce Way
Miami Lakes FL 33014

Direct: +1-607-280-3233
Email: Carlos.Bustamante@galatea.bio
cdbustam@stanford.edu
Assistant: sa.sheng@galatea.bio

Personal Statement

For the past two decades, I have led a multidisciplinary team working on problems in genetics of human health and disease. Much of our research has focused on genomics technology, data science, and its application in medical and basic sciences. My first academic appointment was at Cornell University's College of Agriculture and Life Sciences. There, much of our work focused on population genetics of human health and nutrition motivated by a desire to improve the foods we eat as the basis of health. I moved to Stanford in 2010 to focus on enabling clinical and medical genomics on a global scale. I have been focused on reducing health disparities in genomics broadening representation of understudied groups in large genomic consortia, particularly minority groups from the U.S., the Americas, and Africa. My work has empowered decision-makers to utilize genomics and data science in the service of improving human health and wellbeing. In the next phase of my career, I will focus on opportunities for bringing these technologies to consumer and patients, directly, where this work can have the greatest impact.

I am the Founder, President, and Chief Executive Officer of Galatea Bio, Inc. a company focused on the use of human genomics to power discovery for biopharma. This company was formed while I was a Professor at Stanford and a Venture Partner with Fidelity's early-stage biotechnology and healthcare group, F-Prime Capital Partners. I have a strong interest in building new academic and business units including non-profits and companies. I was the Inaugural Chair of Stanford Medicine's Department of Biomedical Data Science—the first new department that the University had started in 14 years. I serve as an advisor to the US federal government, private companies, startups, and non-profits in the areas of pharmaceutical products, computational genomics, population and medical genetics, intellectual property protection, and business strategy.

Education

Harvard University, Cambridge, MA	B.A.	1997	Biology, <i>magna cum laude</i> with highest honors
Harvard University, Cambridge, MA	M.S.	2001	Statistics
Harvard University, Cambridge, MA	Ph.D.	2001	Biology
University of Oxford, Oxford, UK	Post-doc	2001-2	Mathematical Genetics

Bustamante CV

1

Updated: March 2024

Positions and Employment

04/21-present Galatea Bio, Inc. Founder, President, and CEO
01/24-present Executive-in-Residence, Department of Business Technology University of Miami
Herbert School of Business, Coral Gables, FL
01/19-present F-Prime Capital Partners, Venture Partner
01/16-present Principal and Chairman, CDB Consulting, LTD.
11/22-present Adjunct Professor, Department of Biomedical Data Sciences, Genetics and (by
courtesy) Biology, Stanford University
09/21-06/22 Visiting Lecturer, Department of Business Technology University of Miami Herbert
School of Business, Coral Gables, FL
09/15-11/22 Professor, Department of Biomedical Data Sciences, Stanford University
01/10-11/22 Professor, Department of Genetics and (by courtesy) Biology, Stanford University
03/17-07/22 Investigator, Chan Zuckerberg Initiative (CZI) Biohub
09/15-01/2019 Inaugural Chair, Department of Biomedical Data Science, Stanford University
01/12-09/17 Founding Director, Stanford Center for Computational, Evolutionary, and Human
Genomics
01/12-09/15 Director of Informatics, Stanford Center for Genomics and Personalized Medicine
10/08-01/10 Professor, Department of Biological Statistics and Computational Biology and
Department of Statistical Sciences, Cornell University
08/02-10/08 Assistant Professor, Department of Biological Statistics and Computational Biology
and Department of Statistical Sciences, Cornell University
06/07-01/10 Associate Director, Cornell Center for Comparative and Population Genomics
01/04-06/04 Visiting Scholar, NSF Institute for Pure and Applied Mathematics, Los Angeles

Corporate Leadership

04/21-present Chairman of the Board, Galatea Bio
04/21-present Director, Provaxus, Inc.
01/19-present Board observer, Embark Veterinary, Inc. (Boston, MA)
01/16-present Principal and Chairman, CDB Consulting Group (Miami, FL)
06/18-06/22 Director, Etalon Diagnostics (Palo Alto, CA)
02/15-12/21 Founder, Director, and Chair of Scientific Advisory Board, Arc-Bio, LLC:
formerly Big Data Bio LLC (Boston, MA) and Identify Genomics (Menlo Park, CA)
02/14-06/16 Founder, Director, and Chair of Scientific Advisory Board, Identify Genomics LLC
(Menlo Park, CA)
02/15-12/21 Director, Liberty Biosecurity (Washington, DC)
06/16-12/21 Director, Eden Roc Sciences LLC (Birmingham, MI)
01/19-12/22 Board observer, Genomics PLC (Oxford, UK)

Non-profit and Government Service

06/19-12/22 Chair, ClinGen Complex Disease Working Group
Co-Chair, ClinGen Ancestry and Diversity Working Group
06/15-12/20 Co-Chair, PAGE Analysis Working Group
Chair, ClinGen Informatics Working Group
Member, ClinGen Steering Committee
02/11-12/15 Member, National Human Genome Research Institute (NHGRI) Council
02/11-12/15 National Institutes of Health Council of Councils
01/11-12/14 External Evaluation committee NIDDK T2D GENES project
06/10-12/14 Scientific Advisory Board, SIGMA (Slim Initiative for Genomic Medicine in the
Americas)

11/09-12/20 Scientific Advisory Board, NHGRI Center for Research on Genomics and Global Health

Scientific and Business Advisory Boards (current)

12/23-present Scientific Advisory Board, Colossal Biosciences (Austin, TX)
11/23-present Advisor, Pivotal Biosciences (San Francisco, CA)
06/18-present Scientific and Business Advisory Boards, Awakens, Inc. (Berkeley, CA)
06/18-present Scientific and Business Advisory Boards, Fauna Bio, Inc. (San Francisco, CA)
06/17-present Scientific Advisory Board, LunaDNA, Inc. (San Diego, CA)
06/16-present Scientific Advisory Board, Embark Veterinary (Boston, MA)
06/13-06/22 Scientific Advisory Board, Etalon Diagnostics (Menlo Park, CA)
03/16-12/21 Scientific Advisory Board, Digitalis Ventures (New York, NY)
02/15-12/21 Scientific Advisory Board, Liberty Biosecurity (Washington, DC)
10/17-12/21 Fellow, WebShield, Inc. (San Francisco, CA)
06/18-12/20 Scientific and Business Advisory Boards, Imprimed, Inc. (Palo Alto, CA)
04/18-04/20 Scientific Advisory Board, Columbia Care LLC (New York, NY)
03/18-12/20 Scientific Advisor, Nicklaus Children's Hospital (Miami, FL)
06/11-12/19 Scientific Advisory Board, Ancestry.com (Provo, UT)
06/16-12/18 Mars/IBM Food Safety Board (McClean, VA)
06/13-12/18 Chair of Scientific Advisory Board, Etalon Diagnostics (Menlo Park, CA)
09/11-12/17 Scientific Advisory Board, 23andme "Roots into the Future" (Mountain View, CA)
06/13-12/15 Scientific Advisory Board, National Geographic Genographic Project (Washington, D.C.)
04/12-12/16 Scientific Advisory Board, Personalis, Inc. (Menlo Park, CA)
09/12-12/15 Scientific Advisory Board, InVita (San Francisco, CA)
06/12-12/14 Medical Advisory Board, Med-Tek (Miami, FL)
06/11-06/12 Scientific Advisory Board, Mubadala Healthcare (Abu Dhabi, UAE)
09/09-12/12 Scientific Advisory Board, Nature Source Genetics (Ithaca, NY)

Fellowships and Honors

2017 Chang-Zuckerberg Biohub Investigator
2016 Stanford Prize in Population Genetics and Society
2011-2016 2010 John D. & Catherine T. MacArthur Fellow
2008 Cornell Provost Award for Distinguished Research
2007-2009 Sloan Research Fellowship in Molecular Biology
2001-2002 Marshall-Sherfield Fellowship
1999-2001 Howard Hughes Medical Institute Graduate Research Fellowship
1997-1998 National Science Foundation Graduate Research Fellowship

Editorial Boards

2013-2016 Evolution Section Editor, *PLoS-Genetics*
2013-2016 Editorial Board, *American Journal of Human Genetics*
2010-2020 Editorial Board, *Human Biology*
2008-2017 Editorial Board, *Genome Research*
2008-2012 Editorial Board, *Molecular Biology and Evolution*
2005-2009 Editorial Board, *Bioinformatics*

Other Experience and Professional Memberships

2015-2017	Co-Chair (2015), Chair (2017), Gordon Conference on Human Genetics and Genomics
2012-2014	Co-Chair, Conference on Genomics and Health Disparities
2012-2015	Organizer, Cold Spring Harbor Biology of Genomes Conference
2012-2014	Organizer, Meeting of the Society for Molecular Biology and Evolution
2009-present	Member, The Center for Academic Research and Training in Anthropogeny (UCSD)
2004-2005	Fellow of NSF Institute for Pure and Applied Mathematics

Patents and Patent Applications

1. **US10,981,137 B2. ENRICHMENT OF DNA SEQUENCING LIBRARIES FROM SAMPLES CONTAINING SMALL AMOUNTS OF TARGET DNA.** Priority Date: May 4, 2013. Inventors: **Carlos D. Bustamante**, Meredith Carpenter, Jason Buenrostro, and William J. Greenleaf. **Granted: Apr. 20, 2021.**
2. **US10,954,509 B2. PARTITIONING OF DNA SEQUENCING LIBRARIES INTO HOST AND MICROBIAL COMPONENTS.** Priority Feb 15, 2014. Inventors: **Carlos D. Bustamante** and Meredith Carpenter. **Granted: Mar. 23, 2021.**
3. **US 10,847,248 B2. TECHNIQUES FOR DETERMINING HAPLOTYPE BY POPULATION GENOTYPE AND SEQUENCE DATA.** Priority: Aug. 12, 2013. Inventors: **Carlos D. Bustamante** and Fouad Zakharia. **Granted: Nov. 24, 2020.**
4. **US 10,127,346 B2. SYSTEMS AND METHODS FOR INTERPRETING A HUMAN GENOME USING A SYNTHETIC REFERENCE SEQUENCE.** Priority Date: 2012-04-13. Inventors: Frederick Dewey, Euan A. Ashley, Matthew Wheeler, Michael Snyder, **Carlos Bustamante.** **Granted: Nov. 13, 2018.**
5. **GENOMIC SEQUENCE DATASET GENERATION.** Filing Date: Sep. 14, 2021. Inventors: Daniel Mas Montserrat, Alexander Ioannidis, **Carlos Bustamante.** Pub. No. US20230326542A1.
6. **LOCAL-ANCESTRY INFERENCE WITH MACHINE LEARNING MODEL.** Filing Date: Apr. 15, 2021. Inventors: Daniel Mas Montserrat, Alexander Ioannidis, Arvind Kumar, **Carlos Bustamante**, Richa Rastogi, Helgi Hilmarsson. Pub. No. US20230197204A1.
7. **TECHNIQUES FOR DETERMINING WHETHER AN INDIVIDUAL IS INCLUDED IN ENSEMBLE GENOMIC DATA.** Filing Date: 06.10.2016. Inventors: Suyash Shringarpure **Carlos D. Bustamante.** Pub. No.: WO/2017/062599A1
8. **COMPOSITIONS AND METHODS FOR TARGETED DEPLETION, ENRICHMENT, AND PARTITIONING OF NUCLEIC ACIDS USING CRISPR/CAS SYSTEM PROTEINS.** Filing Date: 19.12.2015. Inventors: Meredith Carpenter, **Carlos D. Bustamante**, Stephane B. Gourguechon. Pub. No.: Wo/2016/100955.
9. **PHASED WHOLE GENOME GENETIC RISK IN A FAMILY QUARTET.** Pub. Date: March 28, 2013. Inventors: Frederick Dewey, Euan A. Ashley, Jake Byrnes, **Carlos D. Bustamante**, Atul J. Butte, Rong Chen. Pub no: 2015/0370959. Grant Number: . **Grant Date: 13.09.2016**
10. **COMPOSITIONS AND METHODS FOR PREDICTING BODY SIZE IN DOGS.** Pub. Date.: 02.11.2007. Inventors: Elaine A. Ostrander, Nathaniel B. Sutter, Pascale Quignon, Robert K. Wayne, Melissa Gray, **Carlos D. Bustamante**, Badri Padhukasahasram, Keyan Zhao, Magnus Nordborg, Ebenezer Satyaraj, Paul G. Jones, Dennis F. Lawler. Pub. No.: WO2008058013A2.

Peer Reviewed Publications (Total 251; 247 peer reviewed)

[star designation (*) denotes I was co-senior author role]

[cross designation (+) denotes I was co-mentor of **lead and/or senior author**]

[Highlighted Note: Describes additional role]

2023

251. Gouveia MH, Bentley AR, Leal TP, Tarazona-Santos E, **Bustamante CD (+)**, Adeyemo AA, Rotimi CN, Shriner D. [Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies](#). **Nat Commun**. 2023 Nov 7;14(1):6802. doi: 10.1038/s41467-023-42491-0. PMID: 37935687; PMCID: PMC10630423.
250. Serrano JG, Ordóñez AC, Santana J, Sánchez-Cañadillas E, Arnay M, Rodríguez-Rodríguez A, Morales J, Velasco-Vázquez J, Alberto-Barroso V, Delgado-Darias T, de Mercadal MCC, Hernández JC, Moreno-Benítez MA, Pais J, Ringbauer H, Sikora M, McColl H, Pino-Yanes M, Ferrer MH, **Bustamante CD*+**, Fregel R*. [The genomic history of the indigenous people of the Canary Islands](#). **Nat Commun**. 2023 Aug 15;14(1):4641. doi: 10.1038/s41467-023-40198-w. PMID: 37582830; PMCID: PMC10427657.

Co-mentor and senior author / group leader for my lab where trainee was senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Rosa Fregel, my former trainee.

249. Mantes AD, Montserrat DM, **Bustamante CD+**, Giró-I-Nieto X, Ioannidis AG+. [Neural ADMIXTURE for rapid genomic clustering](#). **Nat Comput Sci**. 2023 Jul;3(7):621-629. doi: 10.1038/s43588-023-00482-7. Epub 2023 Jul 6. PMID: 37600116; PMCID: PMC10438426.

Co-mentor and senior author / group leader for my lab where trainee was senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Alex Ioannidis, my trainee.

247. Zhou W, Kanai M, Wu KH, Rasheed H, Tsuo K, ..., **Bustamante CD**, ... [Meta-analysis Initiative: Powering genetic discovery across human disease](#). **Cell Genom**. 2022 Oct 12;2(10):100192. doi: 10.1016/j.xgen.2022.100192. PMID: 36777996; PMCID: PMC9903716.

Consortium paper where I was a senior author / group leader for my lab contributing data and methods to the paper. I co-led, financed, and jointly oversaw my groups participation in this project from conception to publication

2022

246. Lopez-Pineda A, Vernekar M, Moreno-Grau S, Rojas-Muñoz A, Moatamed B, Lee MTM, Nava-Aguilar MA, Gonzalez-Arroyo G, Numakura K, Matsuda Y, Ioannidis A, Katsanis N, Takano T, **Bustamante CD**. [Validating and automating learning of cardiometabolic polygenic risk scores from direct-to-consumer genetic and phenotypic data: implications for scaling precision health research](#). **Hum Genomics**. 2022 Sep 8;16(1):37. doi: 10.1186/s40246-022-00406-y. PMID: 36076307

245. Parikh VN, Ioannidis AG, Jimenez-Morales D, Gorzynski JE, De Jong HN, Liu X, Roque J,

Cepeda-Espinoza VP, Osoegawa K, Hughes C, Sutton SC, Youlton N, Joshi R, Amar D, Tanigawa Y, Russo D, Wong J, Lauzon JT, Edelson J, Mas Montserrat D, Kwon Y, Rubinacci S, Delaneau O, Cappello L, Kim J, Shoura MJ, Raja AN, Watson N, Hammond N, Spiteri E, Mallempati KC, Montero-Martín G, Christle J, Kim J, Kirillova A, Seo K, Huang Y, Zhao C, Moreno-Grau S, Hershman SG, Dalton KP, Zhen J, Kamm J, Bhatt KD, Isakova A, Morri M, Ranganath T, Blish CA, Rogers AJ, Nadeau K, Yang S, Blomkalns A, O'Hara R, Neff NF, DeBoever C, Szalma S, Wheeler MT, Gates CM, Farh K, Schroth GP, Febbo P, deSouza F, Cornejo OE, Fernandez-Vina M, Kistler A, Palacios JA, Pinsky BA, **Bustamante CD***, Rivas MA*, Ashley EA*. [Deconvoluting complex correlates of COVID-19 severity with a multi-omic pandemic tracking strategy](#). **Nat Commun**. 2022 Aug 30;13(1):5107. doi: 10.1038/s41467-022-32397-8. PMID: 36042219

244. Gimbernat-Mayol J, Dominguez Mantes A, **Bustamante CD+**, Mas Montserrat D, **Ioannidis AG**. **PLoS Comput Biol**. 2022 Aug 25;18(8):e1010301. doi: 10.1371/journal.pcbi.1010301. [Archetypal Analysis for population genetics](#). eCollection 2022 Aug. PMID: 36007005

Co-mentor and senior author / group leader for my lab where trainee was senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Alex Ioannidis, my trainee.

243. Liu YC, Hunter-Anderson R, Cheronet O, Eakin J, Camacho F, Pietrusewsky M, Rohland N, **Ioannidis A**, Athens JS, Douglas MT, Ikehara-Quebral RM, Bernardos R, Culleton BJ, Mah M, Adamski N, Broomandkhoshbacht N, Callan K, Lawson AM, Mandl K, Michel M, Oppenheimer J, Stewardson K, Zalzal F, Kidd K, Kidd J, Schurr TG, Auckland K, Hill AVS, Mentzer AJ, Quinto-Cortés CD, Robson K, Kennett DJ, Patterson N, **Bustamante CD**, Moreno-Estrada A, Spriggs M, Vilar M, Lipson M, Pinhasi R, Reich D. [Ancient DNA reveals five streams of migration into Micronesia and matrilocality in early Pacific seafarers](#). **Science**. 2022 Jul;377(6601):72-79. doi: 10.1126/science.abm6536. Epub 2022 Jun 30. PMID: 35771911

Consortium paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

242. **Nieves-Colón MA+**, **Badillo Rivera KM+**, **Sandoval K+**, Villanueva Dávalos V, Enriquez Lencinas LE, Mendoza-Revilla J, Adhikari K, González-Buenfil R, Chen JW, Zhang ET, Sockell A, Ortiz-Tello P, Hurtado GM, Condori Salas R, Cebrecos R, Manzaneda Choque JC, Manzaneda Choque FP, Yábar Pilco GP, Rawls E, Eng C, Huntsman S, Burchard E, Ruiz-Linares A, González-José R, Bedoya G, Rothhammer F, Bortolini MC, Poletti G, Gallo C, **Bustamante CD+**, Baker JC*, **Gignoux CR*+**, **Wojcik GL*+**, **Moreno-Estrada A*+**. [Clotting factor genes are associated with preeclampsia in high-altitude pregnant women in the Peruvian Andes](#). **Am J Hum Genet**. 2022 Jun 2;109(6):1117-1139. doi: 10.1016/j.ajhg.2022.04.014. Epub 2022 May 18. PMID: 35588731

Co-mentor and senior author / group leader for my lab where three trainees were co-senior author of the paper. I co-led, co-financed, and jointly oversaw this project from conception to publication with Julie Baker.

241. Preston CG, Wright MW, Madhav Rao R, Harrison SM, Goldstein JL, Luo X, Wand H, Wulf B, Cheung G, Mandell ME, Tong H, Cheng S, Iacocca MA, Pineda AL, Popejoy AB, Dalton K, Zhen J, Dwight SS, Babb L, DiStefano M, O'Daniel JM, Lee K, Riggs ER, Zastrow DB, Mester JL, Ritter DI, Patel RY, Subramanian SL, Milosavljevic A, Berg JS, Rehm HL, Plon SE, Cherry JM, **Bustamante CD***, Costa HA*; Clinical Genome Resource (ClinGen). [ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines](#). **Genome Med**. 2022 Jan 18;14(1):6. doi: 10.1186/s13073-021-01004-8.PMID: 35039090

2021

240. Venkataraman GR, DeBoever C, Tanigawa Y, Aguirre M, **Ioannidis AG+**, Mostafavi H, Spencer CCA, Poterba T, **Bustamante CD+**, Daly MJ, Pirinen M, Rivas MA. [Bayesian model comparison for rare-variant association studies](#). **Am J Hum Genet**. 2021 Dec 2;108(12):2354-2367. doi: 10.1016/j.ajhg.2021.11.005. Epub 2021 Nov 24.PMID: 34822764

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

239. Fu R, Gillen AE, **Grabek KR+**, Riemony KA, Epperson LE, **Bustamante CD*+**, Hesselberth JR, Martin SL*. [Dynamic RNA Regulation in the Brain Underlies Physiological Plasticity in a Hibernating Mammal](#). **Front Physiol**. 2021 Jan 18;11:624677. doi: 10.3389/fphys.2020.624677. eCollection 2020.PMID: 33536943

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Sandy Martin, and my trainee, Katherine Grabek.

238. **Ioannidis AG+**, Blanco-Portillo J, **Sandoval K+**, Hagelberg E, Barberena-Jonas C, Hill AVS, Rodríguez-Rodríguez JE, Fox K, Robson K, Haoa-Cardinali S, Quinto-Cortés CD, Miquel-Poblete JF, Auckland K, Parks T, Sofro ASM, **Ávila-Arcos MC+**, **Sockell A+**, **Homburger JR+**, Eng C, Huntsman S, Burchard EG, **Gignoux CR+**, Verdugo RA, Moraga M, **Bustamante CD*+**, Mentzer AJ, **Moreno-Estrada A*+**. [Paths and timings of the peopling of Polynesia inferred from genomic networks](#). **Nature**. 2021 Sep;597(7877):522-526. doi: 10.1038/s41586-021-03902-8. Epub 2021 Sep 22.PMID: 3455225.

Co-mentor and senior author / group leader for my lab where trainees were first author and senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Alex Ioannidis, my trainee and Andres Moreno-Estrada, my former trainee.

237. **Lopez Pineda A+**, **Pourshafeie A+**, **Ioannidis A+**, Leibold CM, Chan AL, **Bustamante CD+**, Frankovich J, **Wojcik GL+**. [Discovering prescription patterns in pediatric acute-onset neuropsychiatric syndrome patients](#). **J Biomed Inform**. 2021 Jan;113:103664. doi: 10.1016/j.jbi.2020.103664. Epub 2020 Dec 28.PMID: 33359113

Co-mentor and senior author / group leader for my lab where trainees were first author and senior author of the paper. I led, financed, and oversaw this project from conception to publication with G. Wojcik, my trainee at the time the project was initiated.

2020

236. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, **Bustamante CD***, Kenny EE+*. [Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations](#). **Am J Hum Genet**. 2020 Oct 1;107(4):788-789. doi: 10.1016/j.ajhg.2020.08.020. PMID: 33007199

Co-mentor and senior author / group leader for my lab where trainees were first author and senior author of the paper. I led, financed, and oversaw this project from conception to publication with E. Kenny, my trainee at the time the project was initiated.

235. Gorzynski JE, De Jong HN, Amar D, Hughes CR, Ioannidis A, Bierman R, Liu D, Tanigawa Y, Kistler A, Kamm J, Kim J, Cappello L, Neff NF, Rubinacci S, Delaneau O, Shoura MJ, Seo K, Kirillova A, Raja A, Sutton S, Huang C, Sahoo MK, Mallempati KC, Montero-Martin G, Osoegawa K, Jimenez-Morales D, Watson N, Hammond N, Joshi R, Fernandez-Vina M, Christle JW, Wheeler MT, Febbo P, Farh K, Schroth G, Desouza F, Palacios J, Salzman J, Pinsky BA, Rivas MA, **Bustamante CD***, Ashley EA*, Parikh VN*. [High-throughput SARS-CoV-2 and host genome sequencing from single nasopharyngeal swabs](#). medRxiv. 2020 Jul 29:2020.07.27.20163147. doi: 10.1101/2020.07.27.20163147. Preprint. PMID: 32766602

234. Ioannidis AG+, Blanco-Portillo J, Sandoval K+, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, Rodríguez-Rodríguez JE, Quinto-Cortés CD, Auckland K, Parks T, Robson K, Hill AVS, Avila-Arcos MC+, Sockell A+, Homburger JR+, Wojcik GL+, Barnes KC, Herrera L, Berrios S, Acuña M, Llop E, Eng C, Huntsman S, Burchard EG, Gignoux CR+, Cifuentes L, Verdugo RA, Moraga M, Mentzer AJ, **Bustamante CD*+**, Moreno-Estrada A*+. [Native American gene flow into Polynesia predating Easter Island settlement](#). **Nature**. 2020 Jul;583(7817):572-577. doi: 10.1038/s41586-020-2487-2. Epub 2020 Jul 8. PMID: 32641827

Co-mentor and senior author / group leader for my lab where trainees were first author and senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Alex Ioannidis, my trainee and Andres Moreno-Estrada, my former trainee.

233. Venkataraman GR, Pineda AL+, Bear Don't Walk Iv OJ, Zehnder AM, Ayyar S, Page RL, **Bustamante CD***, Rivas MA*. [FasTag: Automatic text classification of unstructured medical narratives](#). **PLoS One**. 2020 Jun 22;15(6):e0234647. doi: 10.1371/journal.pone.0234647. eCollection 2020. PMID: 32569327

Co-mentor and senior author / group leader for my lab where trainee was co-first author. I co-led, financed, and jointly oversaw this project from conception to publication with Manuel Rivas.

232. Popejoy AB+, Crooks KR, Fullerton SM, Hindorff LA, Hooker GW, Koenig BA, Pino N, Ramos EM, Ritter DI, Wand H, Wright MW, Yudell M, Zou JY, Plon SE, **Bustamante CD***, Ormond KE*; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group. [Clinical](#)

[Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures](#). **Am J Hum Genet**. 2020 Jul 2;107(1):72-82. doi: 10.1016/j.ajhg.2020.05.005. Epub 2020 Jun 6. PMID: 32504544

Co-mentor and senior author / group leader for my lab where trainee, Alice Poopojoy, was first author. I co-led, financed, and jointly oversaw this project from conception to publication with Kelly Ormond. I was also PI of the parent grant for ClinGen that financed this study.

231. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, Buyske S; **NHGRI Centers for Common Disease Genomics**, Matise TC, Muzny DM, Zody MC, Lander ES, Dutcher SK, Stitzel NO, Hall IM. [Mapping and characterization of structural variation in 17,795 human genomes](#). **Nature**. 2020 Jul;583(7814):83-89. doi: 10.1038/s41586-020-2371-0. Epub 2020 May 27. PMID: 32460305

Consortium paper where I was a senior author / group leader for my lab contributing data and methods to the paper. I was also MPI of the grant that financed this work.

230. Verdugo RA, Di Genova A, Herrera L, Moraga M, Acuña M, Berríos S, Llop E, Valenzuela CY, Bustamante ML, Digman D, Symon A, Asenjo S, López P, Blanco A, Suazo J, Barozet E, Caba F, Villalón M, Alvarado S, Cáceres D, Salgado K, Portales P, **Moreno-Estrada A+**, **Gignoux CR+**, **Sandoval K+**, **Bustamante CD+**, Eng C, Huntsman S, Burchard EG, Loira N, Maass A, Cifuentes L. [Development of a small panel of SNPs to infer ancestry in Chileans that distinguishes Aymara and Mapuche components](#). **Biol Res**. 2020 Apr 16;53(1):15. doi: 10.1186/s40659-020-00284-5. PMID: 32299502

Consortium paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

229. **ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium**. [Pan-cancer analysis of whole genomes](#). **Nature**. 2020 Feb;578(7793):82-93. doi: 10.1038/s41586-020-1969-6. Epub 2020 Feb 5. PMID: 32025007

Consortium paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

228. Grabek KR, Cooke TF, Epperson LE, Spees KK, Cabral GF, Sutton SC, Merriman DK, Martin SL*, **Bustamante CD***. [Genetic variation drives seasonal onset of hibernation in the 13-lined ground squirrel](#). **Commun Biol**. 2019 Dec 20;2:478. doi: 10.1038/s42003-019-0719-5. eCollection 2019. PMID: 31886416

227. **Ávila-Arcos MC+**, McManus KF, **Sandoval K+**, Rodríguez-Rodríguez JE, Villa-Islas V, Martin AR, Luisi P, Peñaloza-Espinosa RI, Eng C, Huntsman S, Burchard EG, **Gignoux CR**, **Bustamante CD*+**, **Moreno-Estrada A*+**. [Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes](#). **Mol Biol Evol**. 2020 Apr 1;37(4):994-1006. doi: 10.1093/molbev/msz282. PMID: 31848607

Co-mentor and senior author / group leader for my lab where trainees were first author and senior author of the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Maria Avila-Arcos, my trainee, and Chris Gignoux and Andres Moreno-

Estrada, my former trainees.

226. Nie A, Pineda AL, Wright MW, Wand H, Wulf B, Costa HA, Patel RY, **Bustamante CD**, Zou J. [LitGen: Genetic Literature Recommendation Guided by Human Explanations](#). **Pac Symp Biocomput**. 2020; 25:67-78. PMID: 31797587

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

225. **Nieves-Colón MA+**, Pestle WJ, Reynolds AW, Llamas B, de la Fuente C, Fowler K, Skerry KM, Crespo-Torres E, **Bustamante CD***, Stone AC*. [Ancient DNA reconstructs the genetic legacies of pre-contact Puerto Rico communities](#). **Mol Biol Evol**. 2020 Mar 1; 37(3):611-626. doi: 10.1093/molbev/msz267. Epub 2019 Nov 9. PMID: 31710665

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper. I co-led, financed, and jointly oversaw this project from conception to publication with Maria Nieves-Colon, a joint trainee with Anne Stone who was her main supervisor.

2019

224. Musharoff S, Shringarpure S, **Bustamante CD***, Ramachandran S*. [The inference of sex-biased human demography from whole-genome data](#). **PLoS Genet**. 2019 Sep 20;15(9):e1008293. doi: 10.1371/journal.pgen.1008293. eCollection 2019 Sep. PMID: 31539367

223. Lopez M, Choin J, Sikora M, Siddle K, Harmant C, Costa HA, Silvert M, Mouguiama-Daouda P, Hombert JM, Froment A, Le Bomin S, Perry GH, Barreiro LB, **Bustamante CD**, Verdu P, Patin E, Quintana-Murci L. [Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest](#). **Curr Biol**. 2019 Sep 9;29(17):2926-2935.e4. doi: 10.1016/j.cub.2019.07.013. Epub 2019 Aug 8. PMID: 31402299

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper.

222. [Genetic analyses of diverse populations improves discovery for complex traits](#). **Wojcik GL+**, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, ... , **Bustamante CD+**, Caberto C, Canizales-Quinteros S,..., Kooperberg C, Le Marchand L, Loos RJF, Matise TC, North KE, Peters U, Kenny EE, Carlson CS. **Nature**. 2019 Jun;570(7762):514-518. doi: 10.1038/s41586-019-1310-4. Epub 2019 Jun 19. PMID: 31217584

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper. My trainee, Genevieve Wojcik, was first author of the Nature paper on behalf of PAGE consortium. I was also MPI of the grant that financed this work.

221. Bucio D, Ormond KE, Hernandez D, **Bustamante CD*+**, Lopez Pineda A*+. [A genetic counseling needs assessment of Mexico](#). **Mol Genet Genomic Med**. 2019 May;7(5):e668. doi: 10.1002/mgg3.668. Epub 2019 Apr 1. PMID: 30938092

Collaborative paper where I was a senior author / group leader for my lab contributing data

and methods to the paper. My trainee, Aturo Lopez, was senior author.

220. Fregel R, Ordóñez AC, Santana-Cabrera J, Cabrera VM, Velasco-Vázquez J, Alberto V, Moreno-Benítez MA, Delgado-Darias T, Rodríguez-Rodríguez A, Hernández JC, Pais J, González-Montelongo R, Lorenzo-Salazar JM, Flores C, Cruz-de-Mercadal MC, Álvarez-Rodríguez N, Shapiro B, Arnay M, **Bustamante CD**. [Mitogenomes illuminate the origin and migration patterns of the indigenous people of the Canary Islands](#). **PLoS One**. 2019 Mar 20;14(3):e0209125. doi: 10.1371/journal.pone.0209125. eCollection 2019. PMID:30893316
219. **Gignoux CR**+, Torgerson DG, Pino-Yanes M, Uricchio LH, Galanter J, Roth LA, Eng C, Hu D, Nguyen EA, Huntsman S, Mathias RA, Kumar R, Rodriguez-Santana J, Thakur N, Oh SS, McGarry M, **Moreno-Estrada A+**, **Sandoval K**+, Winkler CA, Seibold MA, Padhukasahasram B, Conti DV, Farber HJ, Avila P, Brigino-Buenaventura E, Lenoir M, Meade K, Serebrisky D, Borrell LN, Rodriguez-Cintrón W, Thyne S, Joubert BR, Romieu I, Levin AM, Sienra-Monge JJ, Del Rio-Navarro BE, Gan W, Raby BA, Weiss ST, Bleecker E, Meyers DA, Martinez FJ, Gauderman WJ, Gilliland F, London SJ, **Bustamante CD**+, Nicolae DL, Ober C, Sen S, Barnes K, Williams LK, Hernandez RD, Burchard EG. [An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos](#). **J Allergy Clin Immunol**. 2019 Mar;143(3):957-969. doi: 10.1016/j.jaci.2016.08.057. Epub 2018 Sep 7. PMID: 30201514

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper

218. Li Q, Zhao K, **Bustamante CD**, Ma X, Wong WH. [Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis](#). **Genet Med**. 2019 Sep;21(9):2126-2134. doi: 10.1038/s41436-019-0439-8. Epub 2019 Jan 24. PMID: 3067503

Collaborative paper where I was a senior author / group leader for my lab contributing data and methods to the paper

2018

217. Troll CJ, Putnam NH, Hartley PD, Rice B, Blanchette M, Siddiqui S, Ganbat JO, Powers MP, Ramakrishnan R, Kunder CA, **Bustamante CD**, Zehnder JL, Green RE, Costa HA. [Structural Variation Detection by Proximity Ligation from Formalin-Fixed, Paraffin-Embedded Tumor Tissue](#). **J Mol Diagn**. 2018 Dec 31. pii: S1525-1578(18)30172-7. doi: 10.1016/j.jmoldx.2018.11.003. [Epub ahead of print] PMID: 30605765
215. Lin M, Siford RL, Martin AR, Nakagome S, Möller M, Hoal EG, **Bustamante CD**, Gignoux CR, Henn BM. [Rapid evolution of a skin-lightening allele in southern African KhoeSan](#). **Proc Natl Acad Sci U S A**. 2018 Dec 26;115(52):13324-13329. doi: 10.1073/pnas.1801948115. Epub 2018 Dec 10. PMID: 30530665
214. Huddart R, Fohner AE, Whirl-Carrillo M, Wojcik GL, Gignoux CR, Popejoy AB, **Bustamante CD**, Altman RB, Klein TE. [Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research](#). **Clin Pharmacol Ther**. 2018 Dec 1. doi: 10.1002/cpt.1322. PMID: 30506572

213. Jha AR, Davenport ER, Gautam Y, Bhandari D, Tandukar S, Ng KM, Fragiadakis GK, Holmes S, Gautam GP, Leach J, Sherchand JB, **Bustamante CD**, Sonnenburg JL. [Gut microbiome transition across a lifestyle gradient in Himalaya](#). **PLoS Biol.** 2018 Nov 15;16(11):e2005396. doi: 10.1371/journal.pbio.2005396. eCollection 2018 Nov. PMID: 30439937
212. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H, Wright MW, Zou J, Gignoux CR, Bonham VL, Plon SE, **Bustamante CD**; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG). [The clinical imperative for inclusivity: Race, ethnicity, and ancestry \(REA\) in genomics](#). **Hum Mutat.** 2018 Nov;39(11):1713-1720. doi: 10.1002/humu.23644. PMID:30311373
211. Cornejo OE, Yee MC, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone D 3rd, Stack C, Romero A, Umaharan P, Royaert S, Tawari NR, Ng P, Gutierrez O, Phillips W, Mockaitis K, **Bustamante CD**, Motamayor JC. [Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process](#). **Commun Biol.** 2018 Oct 16;1:167. doi: 10.1038/s42003-018-0168-6. eCollection 2018. PMID:30345393
210. Ioannidis NM, Wang W, Furlotte NA, Hinds DA; 23andMe Research Team, **Bustamante CD**, Jorgenson E, Asgari MM, Whittemore AS. [Gene expression imputation identifies candidate genes and susceptibility loci associated with cutaneous squamous cell carcinoma](#). **Nat Commun.** 2018 Oct 15;9(1):4264. doi: 10.1038/s41467-018-06149-6. PMID:30323283
209. Gignoux CR, Torgerson DG, Pino-Yanes M, Uricchio LH, Galanter J, Roth LA, Eng C, Hu D, Nguyen EA, Huntsman S, Mathias RA, Kumar R, Rodriguez-Santana J, Thakur N, Oh SS, McGarry M, Moreno-Estrada A, Sandoval K, Winkler CA, Seibold MA, Padhukasahasram B, Conti DV, Farber HJ, Avila P, Brigino-Buenaventura E, Lenoir M, Meade K, Serebrisky D, Borrell LN, Rodriguez-Cintron W, Thyne S, Joubert BR, Romieu I, Levin AM, Sienra-Monge JJ, Del Rio-Navarro BE, Gan W, Raby BA, Weiss ST, Bleecker E, Meyers DA, Martinez FJ, Gauderman WJ, Gilliland F, London SJ, **Bustamante CD**, Nicolae DL, Ober C, Sen S, Barnes K, Williams LK, Hernandez RD, Burchard EG. [An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos](#). **J Allergy Clin Immunol.** 2018 Sep 7. pii: S0091-6749(18)31274-0. doi: 10.1016/j.jaci.2016.08.057. PMID: 30201514
208. López-Pineda A, Rodríguez-Moran MF, Álvarez-Aguilar C, Fuentes Valle SM, Acosta-Rosales R, Bhatt AS, Sheth SN, **Bustamante CD**. [Data mining of digitized health records in a resource-constrained setting reveals that timely immunophenotyping is associated with improved breast cancer outcomes](#). **BMC Cancer.** 2018 Sep 27;18(1):933. doi: 10.1186/s12885-018-4833-4. PMID: 30261931
207. Wojcik GL, Fuchsberger C, Taliun D, Welch R, Martin AR, Shringarpure S, Carlson CS, Abecasis G, Kang HM, Boehnke M, **Bustamante CD**, Gignoux CR, Kenny EE. [Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies](#). **G3** (Bethesda). 2018 Aug 21. pii: g3.200502.2018. doi: 10.1534/g3.118.200502. PMID: 30131328
206. Cruz-Dávalos DI, Nieves-Colón MA, Sockell A, Poznik GD, Schroeder H, Stone

- AC, **Bustamante CD**, Malaspinas AS, Ávila-Arcos MC. [In-solution Y-chromosome capture-enrichment on ancient DNA libraries](#). *BMC Genomics*. 2018 Aug 14;19(1):608. doi: 10.1186/s12864-018-4945-x. PMID: 30107783
205. Wang B, Pourshafeie A, Zitnik M, Zhu J, **Bustamante CD**, Batzoglou S, Leskovec J. [Network enhancement as a general method to denoise weighted biological networks](#). *Nat Commun*. 2018 Aug 6;9(1):3108. doi: 10.1038/s41467-018-05469-x. PMID: 30082777
204. Fregel R, Méndez FL, Bokbot Y, Martín-Socas D, Camalich-Massieu MD, Santana J, Morales J, Ávila-Arcos MC, Underhill PA, Shapiro B, Wojcik G, Rasmussen M, Soares AER, Kapp J, Sockell A, Rodríguez-Santos FJ, Mikdad A, Trujillo-Mederos A, **Bustamante CD**. [Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe](#). *Proc Natl Acad Sci U S A*. 2018 Jun 26;115(26):6774-6779. doi: 10.1073/pnas.1800851115. Epub 2018 Jun 12. PMID: 29895688
203. Waszak SM, Northcott PA, Buchhalter I, Robinson GW, Sutter C, Groebner S, Grund KB, Brugières L, Jones DTW, Pajtler KW, Morrissy AS, Kool M, Sturm D, Chavez L, Ernst A, Brabetz S, Hain M, Zichner T, Segura-Wang M, Weischenfeldt J, Rausch T, Mardin BR, Zhou X, Baciú C, Lawrenz C, Chan JA, Varlet P, Guerrini-Rousseau L, Fults DW, Grajkowska W, Hauser P, Jabado N, Ra YS, Zitterbart K, Shringarpure SS, De La Vega FM, **Bustamante CD**, Ng HK, Perry A, MacDonald TJ, Hernáiz Driever P, Bendel AE, Bowers DC, McCowage G, Chintagumpala MM, Cohn R, Hassall T, Fleischhack G, Eggen T, Wesenberg F, Feychting M, Lannering B, Schüz J, Johansen C, Andersen TV, Rösli M, Kuehni CE, Grotzer M, Kjaerheim K, Monoranu CM, Archer TC, Duke E, Pomeroy SL, Shelagh R, Frank S, Sumerauer D, Scheurlen W, Ryzhova MV, Milde T, Kratz CP, Samuel D, Zhang J, Solomon DA, Marra M, Eils R, Bartram CR, von Hoff K, Rutkowski S, Ramaswamy V, Gilbertson RJ, Korshunov A, Taylor MD, Lichter P, Malkin D, Gajjar A, Korbel JO, Pfister SM. [Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort](#). *Lancet Oncol*. 2018 Jun;19(6):785-798. doi: 10.1016/S1470-2045(18)30242-0. Epub 2018 May 9. PMID: 29753700
202. Muzzio M, Motti JMB, Paz Sepulveda PB, Yee MC, Cooke T, Santos MR, Ramallo V, Alfaro EL, Dipierri JE, Bailliet G, Bravi CM, **Bustamante CD**, Kenny EE. [Population structure in Argentina](#). *PLoS One*. 2018 May 1;13(5):e0196325. doi: 10.1371/journal.pone.0196325. eCollection 2018. PMID: 29715266
201. DeBoever C, Tanigawa Y, Lindholm ME, McInnes G, Lavertu A, Ingelsson E, Chang C, Ashley EA, **Bustamante CD**, Daly MJ, Rivas MA. [Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study](#). *Nat Commun*. 2018 Apr 24;9(1):1612. doi: 10.1038/s41467-018-03910-9. PMID: 29691392
200. de Freitas PD, Mendez FL, Chávez-Congrains K, Galetti PM Jr, Coutinho LL, Pissinatti A, **Bustamante CD**. [Next-Generation Sequencing of the Complete Mitochondrial Genome of the Endangered Species Black Lion Tamarin *Leontopithecus chrysopygus* \(Primates\) and Mitogenomic Phylogeny Focusing on the Callitrichidae Family](#). *G3* (Bethesda). 2018 May 31;8(6):1985-1991. doi: 10.1534/g3.118.200153. PMID: 29650540
199. de la Fuente C, Ávila-Arcos MC, Galimany J, Carpenter ML, Homburger JR, Blanco A,

Contreras P, Cruz Dávalos D, Reyes O, San Roman M, Moreno-Estrada A, Campos PF, Eng C, Huntsman S, Burchard EG, Malaspinas AS, **Bustamante CD**, Willerslev E, Llop E, Verdugo RA, Moraga M. [Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia](#). **Proc Natl Acad Sci U S A**. 2018 Apr 24;115(17):E4006-E4012. doi: 10.1073/pnas.1715688115. Epub 2018 Apr 9. PMID: 29632188

198. Bhattacharya S, Li J, Sockell A, Kan MJ, Bava FA, Chen SC, Ávila-Arcos MC, Ji X, Smith E, Asadi NB, Lachman RS, Lam HYK, **Bustamante CD**, Butte AJ, Nolan GP. [Whole-genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia](#). **Genome Res**. 2018 Apr;28(4):423-431. doi: 10.1101/gr.223693.117. Epub 2018 Mar 22. PMID: 29567674
197. Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Maisano Delser P, Sandoval Velasco M, Schraiber JG, Rasmussen S, Homburger JR, Ávila-Arcos MC, Allentoft ME, Moreno-Mayar JV, Renaud G, Gómez-Carballea A, Laffoon JE, Hopkins RJA, Higham TFG, Carr RS, Schaffer WC, Day JS, Hoogland M, Salas A, **Bustamante CD**, Nielsen R, Bradley DG, Hofman CL, Willerslev E. [Origins and genetic legacies of the Caribbean Taino](#). **Proc Natl Acad Sci U S A**. 2018 Mar 6;115(10):2341-2346. doi: 10.1073/pnas.1716839115. Epub 2018 Feb 20. PMID: 29463742
196. Rappoport N, Toung J, Hadley D, Wong RJ, Fujioka K, Reuter J, Abbott CW, Oh S, Hu D, Eng C, Huntsman S, Bodian DL, Niederhuber JE, Hong X, Zhang G, Sikora-Wohfeld W, Gignoux CR, Wang H, Oehlert J, Jelliffe-Pawlowski LL, Gould JB, Darmstadt GL, Wang X, **Bustamante CD**, Snyder MP, Ziv E, Patsopoulos NA, Muglia LJ, Burchard E, Shaw GM, O'Brodovich HM, Stevenson DK, Butte AJ, Sirota M. [A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth](#). **Sci Rep**. 2018 Jan 9;8(1):226. doi: 10.1038/s41598-017-18246-5. PMID: 29317701
195. Lins LSF, Trojahn S, Sockell A, Yee MC, Tatarenkov A, **Bustamante CD**, Earley RL, Kelley JL. [Whole-genome sequencing reveals the extent of heterozygosity in a preferentially self-fertilizing hermaphroditic vertebrate](#). **Genome**. 2018 Apr;61(4):241-247. doi: 10.1139/gen-2017-0188. Epub 2017 Dec 21. PMID: 29268023
194. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, Atkinson EG, Werely CJ, Möller M, Sandhu MS, Kingsley DM, Hoal EG, Liu X, Daly MJ, Feldman MW, Gignoux CR, **Bustamante CD**, Henn BM. [An Unexpectedly Complex Architecture for Skin Pigmentation in Africans](#). **Cell**. 2017 Nov 30;171(6):1340-1353.e14. doi: 10.1016/j.cell.2017.11.015. PMID: 29195075

2017

193. GTEx Consortium. [Genetic effects on gene expression across human tissues](#). **Nature**. 2017. 550(7675):204-213. doi: 10.1038/nature24277. PMID: 29022597
192. Cooke TF, Fischer CR, Wu P, Jiang TX, Xie KT, Kuo J, Doctorov E, Zehnder A, Khosla C, Chuong CM, **Bustamante CD**. [Genetic Mapping and Biochemical Basis of Yellow Feather Pigmentation in Budgerigars](#). **Cell**. 2017. 171(2):427-439.e21.doi:10.1016/j.cell.2017.08.016. PMID: 28985565
191. Ioannidis NM, Davis JR, DeGorter MK, Larson NB, McDonnell SK, French AJ, Battle AJ, Hastie TJ, Thibodeau SN, Montgomery SB, **Bustamante CD**, Sieh W, Whittemore AS. [FIRE:](#)

[functional inference of genetic variants that regulate gene expression](#). **Bioinformatics**. 2017. doi: 10.1093/bioinformatics/btx534. PMID: 28961785

190. Migicovsky Z, Sawler J, Gardner KM, Aradhya MK, Prins BH, Schwaninger HR, **Bustamante CD**, Buckler ES, Zhong GY, Brown PJ, Myles S. [Patterns of genomic and phenomic diversity in wine and table grapes](#). **Hortic Res**. 2017. 4:17035. doi: 10.1038/hortres.2017.35. eCollection 2017. PMID: 28791127
189. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, Heyer E, Massougbodji A, Fortes-Lima C, Migot-Nabias F, Bellis G, Dugoujon JM, Pereira JB, Fernandes V, Pereira L, Van der Veen L, Mouguiama-Daouda P, **Bustamante CD**, Hombert JM, Quintana-Murci L. [Dispersals and genetic adaption of Bantu-speaking populations in Africa and North America](#). **Science**. 2017. 356(6337):543-546. doi: 10.1126/science.aal1988. PMID: 28473590
188. Holl HM, Brooks SA, Carpenter ML, Bustamante CD, Lafayette C. [A novel splice mutation within equine KIT and the W15 allele in the homozygous state lead to all white coat color phenotypes](#). **Anim Genet**. 2017. 48(4):497-498. doi: 10.1111/age.12554. Epub 2017 Apr 5. PMID: 28378922
187. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, **Bustamante CD**, Kenny EE. [Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations](#). **Am J Hum Genet**. 2017. 100(4):635-649. Doi: 10.1016/j.ajhg.2017.03.004. Epub 2017 Mar 30. PMID: 28366442
186. Costa HA, Neal JW, Bustamante CD, Zehnder JL. [Identification of a Novel Somatic Mutation Leading to Allele Dropout for EGFR L858R Genotyping in Non-Small Cell Lung Cancer](#). **Mol Diagn Ther**. 2017. (4):431-436. doi: 10.1007/s40291-017-0275-y. PMID: 28357677
185. McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE. [Population genetic analysis of the DARC locus \(Duffy\) reveals adaptation from standing variation associated with malaria resistance in humans](#). **PLoS Genet**. 2017 13(3):e1006560. doi: 10.1371/journal.pgen.1006560. eCollection 2017 Mar. PMID:28282382
184. Shringarpure SS, Mathias RA, Hernandez RD, O'Connor TD, Szpiech ZA, Torres R, De La Vega FM, **Bustamante CD**, Barnes KC, Taub MA; CAAPA Consortium. 2016. [Using genotype array data to compare multi- and single-sample variant calls and improve variant call sets from deep coverage whole-genome sequencing data](#). **Bioinformatics**. 2017. 33(8):1147-1153. doi:10.1093/bioinformatics/btw786. PMID: 28035032.

2016

183. Bien SA, Wojcik GL, Zubair N, Gignoux CR, Martin AR, Kocarnik JM, Martin LW, Buyske S, Haessler J, Walker RW, Cheng I, Graff M, Xia L, Franceschini N, Matise T, James R, Hindorff L, Le Marchand L, North KE, Haiman CA, Peters U, Loos RJ, Kooperberg CL, **Bustamante CD**, Kenny EE, Carlson CS; PAGE Study. 2016. [Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array](#). **PLoS One**. 2016. 11(12):e0167758. doi: 10.1371/journal.pone.0167758. PMID: 27973554.
182. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, Vergara C, Torgerson

DG, Pino-Yanes M, Shringarpure SS, Huang L, Rafaels N, Boorgula MP, Johnston HR, Ortega VE, Levin AM, Song W, Torres R, Padhukasahasram B, Eng C, Mejia-Mejia DA, Ferguson T, Qin ZS, Scott AF, Yazdanbakhsh M, Wilson JG, Marrugo J, Lange LA, Kumar R, Avila PC, Williams LK, Watson H, Ware LB, Olopade C, Olopade O, Oliveira R, Ober C, Nicolae DL, Meyers D, Mayorga A, Knight-Madden J, Hartert T, Hansel NN, Foreman MG, Ford JG, Faruque MU, Dunston GM, Caraballo L, Burchard EG, Bleecker E, Araujo MI, Herrera-Paz EF, Gietzen K, Grus WE, Bamshad M, **Bustamante CD**, Kenny EE, Hernandez RD, Beaty TH, Ruczinski I, Akey J; CAAPA., Barnes KC. 2016. [A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome](#). **Nat Comm** 7:12522. doi: 10.1038/ncomms12522. PMID: 27725671.

181. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, **Bustamante CD**, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. 2016. [REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants](#). **Am J Hum Genet**. 2016 Oct 6;99(4):877-885. doi: 10.1016/j.ajhg.2016.08.016. PMID: 27666373.
180. Daneshjou R, Cavallari LH, Weeke PE, Karczewski KJ, Drozda K, Perera MA, Johnson JA, Klein TE, **Bustamante CD**, Roden DM, Shaffer C, Denny JC, Zehnder JL, Altman RB. 2016. [Population-specific single-nucleotide polymorphism confers increased risk of venous thromboembolism in African Americans](#). **Mol Genet Genomic Med**. 2016 Jun 21;4(5):513-20. doi: 10.1002/mgg3.226. PMID: 27652279.
179. Kelley JL, Yee MC, Brown AP, Richardson RR, Tatarenkov A, Lee CC, Harkins TT, **Bustamante CD**, Earley RL. 2016. [The Genome of the Self-Fertilizing Mangrove Rivulus Fish, *Kryptolebias marmoratus*: A Model for Studying Phenotypic Plasticity and Adaptations to Extreme Environments](#). **Genome Biol Evol**. 2016 Aug 16;8(7):2145-54. doi: 10.1093/gbe/evw145. PMID: 27324916.
178. Homburger JR, Green EM, Caleshu C, Sunitha MS, Taylor RE, Ruppel KM, Metpally RP, Colan SD, Michels M, Day SM, Olivotto I, **Bustamante CD**, Dewey FE, Ho CY, Spudich JA, Ashley EA. 2016. [Multidimensional structure-function relationships in human \$\beta\$ -cardiac myosin from population-scale genetic variation](#). **Proc Natl Acad Sci U S A**. 2016 Jun 14;113(24):6701-6. doi: 10.1073/pnas.1606950113. PMID: 27247418.
177. Mendez FL, Poznik GD, Castellano S, **Bustamante CD**. 2016. [The Divergence of Neandertal and Modern Human Y Chromosomes](#). **Am J Hum Genet** 98(4):728-34. doi: 10.1016/j.ajhg.2016.02.023. PMID: 27058445.
176. Altman RB, Prabhu S, Sidow A, Zook JM, Goldfeder R, Litwack D, Ashley E, Asimenos G, **Bustamante CD**, Donigan K, Giacomini KM, Johansen E, Khuri N, Lee E, Liang XS, Salit M, Serang O, Tezak Z, Wall DP, Mansfield E, Kass-Hout T. 2016. [A research roadmap for next-generation sequencing informatics](#). **Sci Transl Med** 8(335):335ps10. doi: 10.1126/scitranslmed.aaf7314. Review. PMID: 27099173.
175. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X,

- Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE; 1000 Genomes Project Consortium, **Bustamante CD**, Tyler-Smith C. 2016. [Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences](#). **Nat Genet** 48(6):593-9. doi: 10.1038/ng.3559. Epub 2016 Apr 25. PMID: 27111036.
174. Shringarpure SS, **Bustamante CD**, Lange K, Alexander DH. 2016. [Efficient analysis of large datasets and sex bias with ADMIXTURE](#). **BMC Bioinformatics** 17:218. doi: 10.1186/s12859-016-1082-x. PMID: 27216439.
173. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, **Bustamante CD**, Kenny EE, Williams SM, Aldrich MC, Gravel S. 2016. [The Great Migration and African-American Genomic Diversity](#). **PLoS Genet** 12(5):e1006059. doi: 10.1371/journal.pgen.1006059. eCollection 2016 May. PMID: 27232753.
172. Homburger JR, Green EM, Caleshu C, Sunitha MS, Taylor RE, Ruppel KM, Metpally RP, Colan SD, Michels M, Day SM, Olivotto I, **Bustamante CD**, Dewey FE, Ho CY, Spudich JA, Ashley EA. 2016. [Multidimensional structure-function relationships in human \$\beta\$ -cardiac myosin from population-scale genetic variation](#). **Proc Natl Acad Sci U S A** 113(24):6701-6. doi: 10.1073/pnas.1606950113. Epub 2016 May 31. PMID: 27247418.
171. Kelley JL, Arias-Rodriguez L, Martin DP, Yee MC, **Bustamante CD**, Tobler M. 2016. [Mechanisms underlying adaptation to life in hydrogen sulfide rich environments](#). **Mol Biol Evol**. 2016 Jun 19. pii: evw145. PMID: 26861137.
170. Davis JR, Fresard L, Knowles DA, Pala M, **Bustamante CD**, Battle A, Montgomery SB. 2016. [An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants](#). **Am J Hum Genet**. 98(1):216-24. PMID: 26749306.
169. Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, Kelley JL, Bailliet G, Bravi CM, **Bustamante CD**, Kenny EE. 2016. [GBStools: A Statistical Method for Estimating Allelic Dropout in Reduced Representation Sequencing Data](#). **PLoS Genet**. 12(2):e1005631. PMID: 26828719.
168. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, Excoffier L, Kidd JM, **Bustamante CD**. 2016. [Distance from sub-Saharan Africa predicts mutational load in diverse human genomes](#). **Proc Natl Acad Sci U S A**. 113(4):E440-9. PMID: 26712023.
167. Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadiogtla V, Siepel A, **Bustamante CD**, Harkins TT, Nelson SF, Marques-Bonet T, Ostrander EA, Wayne RK, Novembre J. 2016. [Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs](#). **PLoS Genet**. 2016 Mar 4;12(3):e1005851. doi: 10.1371/journal.pgen.1005851.

2015

166. Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF; Great Ape Genome Project, **Bustamante CD**, Hammer MF, Wall JD. 2015. [The Time Scale of Recombination Rate Evolution in Great Apes](#). **Mol Biol Evol**. 2015 Dec 15. pii: msv331. PMID: 26671457.
165. Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, Clément-Ziza M, Willemsen D, Cui R, Harel I, Machado BE, Yee MC, Sharp SC, **Bustamante CD**, Beyer A, Johnson EA, Brunet A. 2015. [The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan](#). **Cell**. 163(6):1539-54. PMID: 26638078.
164. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, Gravel S, Alarcón-Riquelme ME, **Bustamante CD** 2015. [Genomic Insights into the Ancestry and Demographic History of South America](#). **PLoS Genet**. 11(12):e1005602. PMID: 26636962.
163. Costa HA, Leitner MG, Sos ML, Mavrantoni A, Rychkova A, Johnson JR, Newton BW, Yee MC, De La Vega FM, Ford JM, Krogan NJ, Shokat KM, Oliver D, Halaszovich CR, **Bustamante CD**. 2015. [Discovery and functional characterization of a neomorphic PTEN mutation](#). **Proc Natl Acad Sci U S A**. 112(45):13976-81. PMID: 26504226.
162. Shringarpure SS, **Bustamante CD**. 2015. [Privacy Risks from Genomic Data-Sharing Beacons](#). **Am J Hum Genet**. 97(5):631-46. PMID: 26522470.
161. Shannon LM, Boyko RH, Castelhamo M, Corey E, Hayward JJ, McLean C, White ME, Abi Said M, Anita BA, Bondjengo NI, Calero J, Galov A, Hedimbi M, Imam B, Khalap R, Lally D, Masta A, Oliveira KC, Pérez L, Randall J, Tam NM, Trujillo-Cornejo FJ, Valeriano C, Sutter NB, Todhunter RJ, **Bustamante CD**, Boyko AR. 2015. [Genetic structure in village dogs reveals a Central Asian domestication origin](#). **Proc Natl Acad Sci U S A**. 112(44):13639-44. PMID: 26483491.
160. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A, Heidari N, Euskirchen G, Huber W, Pritchard JK, **Bustamante CD**, Steinmetz LM, Kundaje A, Snyder M. 2015. [Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions](#). **Cell** 162(5):1051-65. PMID: 26300125.
159. Patwardhan A, Harris J, Leng N, Bartha G, Church DM, Luo S, Haudenschild C, Pratt M, Zook J, Salit M, Tirch J, Morra M, Chervitz S, Li M, Clark M, Garcia S, Chandratillake G, Kirk S, Ashley E, Snyder M, Altman R, **Bustamante C**, Butte AJ, West J, Chen R. 2015. [Achieving high-sensitivity for clinical applications using augmented exome sequencing](#). **Genome Med** 7(1):71. PMID: 26269718.
158. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas AS, Eriksson A, Moltke I, Metspalu M, Homburger JR, Wall J, Cornejo OE, Moreno-Mayar JV, Korneliussen TS, Pierre T, Rasmussen M, Campos PF, Damgaard PB, Allentoft ME, Lindo J, Metspalu E, Rodríguez-

Varela R, Mansilla J, Henrickson C, Seguin-Orlando A, Malmström H, Stafford T Jr, Shringarpure SS, Moreno-Estrada A, Karmin M, Tambets K, Bergström A, Xue Y, Warmuth V, Friend AD, Singarayer J, Valdes P, Balloux F, Lebreiro I, Vera JL, Rangel-Villalobos H, Pettener D, Luiselli D, Davis LG, Heyer E, Zollikofer CP, Ponce de León MS, Smith CI, Grimes V, Pike KA, Deal M, Fuller BT, Arriaza B, Standen V, Luz MF, Ricaut F, Guidon N, Osipova L, Voevoda MI, Posukh OL, Balanovsky O, Lavryashina M, Bogunov Y, Khusnutdinova E, Gubina M, Balanovska E, Fedorova S, Litvinov S, Malyarchuk B, Derenko M, Mosher MJ, Archer D, Cybulski J, Petzelt B, Mitchell J, Worl R, Norman PJ, Parham P, Kemp BM, Kivisild T, Tyler-Smith C, Sandhu MS, Crawford M, Villems R, Smith DG, Waters MR, Goebel T, Johnson JR, Malhi RS, Jakobsson M, Meltzer DJ, Manica A, Durbin R, **Bustamante CD**, Song YS, Nielsen R, Willerslev E. 2015. [Genomic evidence for the Pleistocene and recent population history of Native Americans](#). **Science**. 2015 Jul 21. pii: aab3884. PMID: 26198033.

157. Shringarpure SS, Carroll A, De La Vega FM, **Bustamante CD**. 2015. [Inexpensive and Highly Reproducible Cloud-Based Variant Calling of 2,535 Human Genomes](#). 2015. **PLoS One**. 2015 Jun 25;10(6):e0129277. doi: 10.1371/journal.pone.0129277. eCollection 2015. PMID: 26110529.
156. Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CP, Ponce de León MS, Allentoft ME, Moltke I, Jónsson H, Valdiosera C, Malhi RS, Orlando L, **Bustamante CD**, Stafford TW Jr, Meltzer DJ, Nielsen R, Willerslev E. 2015. [The ancestry and affiliations of Kennewick Man](#). **Nature**. 2015 Jul 23;523(7561):455-8. doi: 10.1038/nature14625. PMID: 26087396.
154. Rehm HL, Berg JS, Brooks LD, **Bustamante CD**, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS; ClinGen. 2015. [ClinGen--the Clinical Genome Resource](#). **N Engl J Med**. 2015 Jun 4;372(23):2235-42. doi: 10.1056/NEJMs1406261. Epub 2015 May 27. No abstract available. PMID: 26014595.
153. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, **Bustamante CD**, Rasmussen M, Snyder MP. 2015. [Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events](#). **Nat Biotechnol**. 2015 Jul;33(7):736-42. doi: 10.1038/nbt.3242. Epub 2015 May 18. PMID: 25985263.
152. Henn BM, Botigué LR, **Bustamante CD**, Clark AG, Gravel S. 2015. [Estimating the mutation load in human genomes](#). **Nat Rev Genet**. 2015 Jun;16(6):333-43. doi: 10.1038/nrg3931. Epub 2015 May 12. Review. PMID: 25963372.
151. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, Ferreira PG, Smith KS, Zhang R, Zhao F, Banks E, Poplin R, Ruderfer DM, Purcell SM, Tukiainen T, Minikel EV, Stenson PD, Cooper DN, Huang KH, Sullivan TJ, Nedzel J; GTEx Consortium; Geuvadis Consortium, **Bustamante CD**, Li JB, Daly MJ, Guigo R, Donnelly P, Ardlie K, Sammeth M, Dermitzakis ET, McCarthy MI, Montgomery SB, Lappalainen T, MacArthur DG. 2015. [Human genomics. Effect of predicted](#)

[protein-truncating genetic variants on the human transcriptome](#). **Science**. 2015 May 8;348(6235):666-9. PMID: 25954003.

150. Schroeder H, Ávila-Arcos MC, Malaspinas AS, Poznik GD, Sandoval-Velasco M, Carpenter ML, Moreno-Mayar JV, Sikora M, Johnson PL, Allentoft ME, Samaniego JA, Haviser JB, Dee MW, Stafford TW Jr, Salas A, Orlando L, Willerslev E, **Bustamante CD**, Gilbert MT. [Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean](#). Proc Natl Acad Sci U S A. 2015 Mar 24;112(12):3669-73. doi: 10.1073/pnas.1421784112. Epub 2015 Mar 9. PMID: 25755263.
149. McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Ape Genome Project G, Kidd JM, Wall JD, **Bustamante CD**, Hammer MF. [Inference of gorilla demographic and selective history from whole-genome sequence data](#). Mol Biol Evol. 2015 Mar;32(3):600-12. doi: 10.1093/molbev/msu394. Epub 2014 Dec 21. PMID: 25534031.

2014

148. Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, Huntsman S, Nishimura KK, Gourraud PA, Mohajeri K, O'Roak BJ, Hu D, Mathias RA, Nguyen EA, Roth LA, Padhukasahasram B, Moreno-Estrada A, Sandoval K, Winkler CA, Lurmann F, Davis A, Farber HJ, Meade K, Avila PC, Serebrisky D, Chapela R, Ford JG, Lenoir MA, Thyne SM, Brigino-Buenaventura E, Borrell LN, Rodriguez-Cintron W, Sen S, Kumar R, Rodriguez-Santana JR, **Bustamante CD**, Martinez FD, Raby BA, Weiss ST, Nicolae DL, Ober C, Meyers DA, Bleecker ER, Mack SJ, Hernandez RD, Eichler EE, Barnes KC, Williams LK, Torgerson DG, Burchard EG. [Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos](#). 2015. **J Allergy Clin Immunol** pii: S0091-6749(14)01575-9. PMID: 25488688.
147. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Grønnow B, Appelt M, Gulløv HC, Friesen TM, Fitzhugh W, Malmström H, Rasmussen S, Olsen J, Melchior L, Fuller BT, Fahrni SM, Stafford T Jr, Grimes V, Renouf MA, Cybulski J, Lynnerup N, Lahr MM, Britton K, Knecht R, Arneborg J, Metspalu M, Cornejo OE, Malaspinas AS, Wang Y, Rasmussen M, Raghavan V, Hansen TV, Khusnutdinova E, Pierre T, Dneprovsky K, Andreasen C, Lange H, Hayes MG, Coltrain J, Spitsyn VA, Götherström A, Orlando L, Kivisild T, Vilems R, Crawford MH, Nielsen FC, Dissing J, Heinemeier J, Meldgaard M, **Bustamante CD**, O'Rourke DH, Jakobsson M, Gilbert MT, Nielsen R, Willerslev E. [The genetic prehistory of the New World Arctic](#). 2014. **Science** 345(6200):1255832. PMID: 25170159.
146. Daneshjou R, Gamazon ER, Burkley B, Cavallari LH, Johnson JA, Klein TE, Limdi N, Hillenmeyer S, Percha B, Karczewski KJ, Langae T, Patel SR, **Bustamante CD**, Altman RB, Perera MA. [Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans](#). **Blood** 124(14):2298-305. PMID: 25079360.
145. Pino-Yanes M, Thakur N, Gignoux CR, Galanter JM, Roth LA, Eng C, Nishimura KK, Oh SS, Vora H, Huntsman S, Nguyen EA, Hu D, Drake KA, Conti DV, Moreno-Estrada A, Sandoval K, Winkler CA, Borrell LN, Lurmann F, Islam TS, Davis A, Farber HJ, Meade K, Avila PC, Serebrisky D, Bibbins-Domingo K, Lenoir MA, Ford JG, Brigino-Buenaventura E, Rodriguez-Cintron W, Thyne SM, Sen S, Rodriguez-Santana JR, **Bustamante CD**, Williams LK, Gilliland FD, Gauderman WJ, Kumar R, Torgerson DG, Burchard EG. [Genetic ancestry influences asthma susceptibility and lung function among Latinos](#). 2014. **J Allergy Clin Immunol**. S0091-6749(14)01107-5. PMID: 25301036.

144. Kelley JL, Peyton JT, Fiston-Lavier AS, Teets NM, Yee MC, Johnston JS, **Bustamante CD**, Lee RE, Denlinger DL. [Compact genome of the Antarctic midge is likely an adaptation to an extreme environment.](#) 2014. **Nat Commun** 5:4611. PMID: 25118180.
143. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, Grubert F, Cann HM, Snyder M, Montgomery SB, **Bustamante CD**. [Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture.](#) 2014. **PLoS Genet** 10(8):e1004549. PMID: 25121757.
142. Gardner KM, Brown PJ, Cooke TF, Cann S, Costa F, **Bustamante CD**, Velasco R, Troggio M, Myles S. [Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing.](#) **G3**. 2014 Jul 16. pii: g3.114.011023. doi: 10.1534/g3.114.011023. PMID: 25031181.
141. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-Alonso V, Sandoval K, Eng C, Romero-Hidalgo S, Ortiz-Tello P, Robles V, Kenny EE, Nuño-Arana I, Barquera-Lozano R, Macín-Pérez G, Granados-Arriola J, Huntsman S, Galanter JM, Via M, Ford JG, Chapela R, Rodríguez-Cintrón W, Rodríguez-Santana JR, Romieu I, Sienra-Monge JJ, del Rio Navarro B, London SJ, Ruiz-Linares A, Garcia-Herrera R, Estrada K, Hidalgo-Miranda A, Jimenez-Sanchez G, Carnevale A, Soberón X, Canizales-Quinteros S, Rangel-Villalobos H, Silva-Zolezzi I, Burchard EG, **Bustamante CD**. [The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.](#) **Science**. 2014. 344:1280-5. doi: 10.1126/science.1251688. PMID: 24926019.
140. Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, Zara I, Pitzalis M, Sidore C, Busonero F, Maschio A, Angius A, Jones C, Mendoza-Revilla J, Nekhrizov G, Dimitrova D, Theodossiev N, Harkins TT, Keller A, Maixner F, Zink A, Abecasis G, Sanna S, Cucca F, **Bustamante CD**. [Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe.](#) 2014. **PLoS Genet** 10(5):e1004353. PMID: 24809476.
139. Kidd JM, Shapton TJ, Bobo D, Norman PJ, Martin AR, Carpenter ML, Sikora M, Gignoux CR, Nemat-Gorgani N, Adams A, Guadalupe M, Guo X, Feng Q, Li Y, Liu X, Parham P, Hoal EG, Feldman MW, Pollard KS, Wall JD, Bustamante CD, Henn BM. [Exome capture from saliva produces high quality genomic and metagenomic data.](#) 2013. **BMC Genomics** 15(1):262. PMID: 2470809.
138. Underhill PA, Poznik GD, Rootsi S, Järve M, Lin AA, Wang J, Passarelli B, Kanbar J, Myres NM, King RJ, Di Cristofaro J, Sahakyan H, Behar DM, Kushniarevich A, Sarac J, Saric T, Rudan P, Pathak AK, Chaubey G, Grugni V, Semino O, Yepiskoposyan L, Bahmanimehr A, Farjadian S, Balanovsky O, Khusnutdinova EK, Herrera RJ, Chiaroni J, Bustamante CD, Quake SR, Kivisild T, Villems R. [The phylogenetic and geographic structure of Y-chromosome haplogroup R1a.](#) 2014. **Eur J Hum Genet.** (in press). doi: 10.1038/ejhg.2014.50. PMID: 24667786.
137. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, Poznik GD, ..., **Bustamante CD**, Jakobsson M, Nielsen R, Willerslev E. [The genome of a Late Pleistocene human from a Clovis burial site in western Montana.](#) **Nature**. 2014 Feb 13;506(7487):225-9. doi: 10.1038/nature13025. PMID: 24522598.
136. Drake KA, Torgerson DG, Gignoux CR, Galanter JM, Roth LA, Huntsman S, Eng C, Oh SS, Yee SW, Lin L, Bustamante CD, Moreno-Estrada A, Sandoval K, Davis A, Borrell LN, Farber HJ, Kumar R, Avila PC, Brigino-Buenaventura E, Chapela R, Ford JG, Lenoir MA, Lurmann F, Meade K, Serebrisky D, Thyne S, Rodríguez-Cintrón W, Sen S, Rodríguez-Santana JR, Hernandez RD, Giacomini KM, Burchard EG. [A genome-wide association study of bronchodilator response in Latinos implicates rare variants.](#) 2014. **J Allergy Clin Immunol.** 133(2):370-378.e15. doi: 10.1016/j.jaci.2013.06.043. PMID:

23992748.

135. Martin AR, Tse G, **Bustamante CD**, Kenny EE. [*Imputation-based assessment of next generation rare exome variant arrays*](#). **Pac Symp Biocomput**. 2014:241-52. PMID: 24297551.
134. Daneshjou R, Zappala Z, Kukurba K, Boyle SM, Ormond KE, Klein TE, Snyder M, **Bustamante CD**, Altman RB, Montgomery SB. [*Path-scan: a reporting tool for identifying clinically actionable variants*](#). 2014. **Pac Symp Biocomput** 2014:229-40. PMID: 24297550.
133. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadiogola V, Siepel A, **Bustamante CD**, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK, Novembre J. [*Genome sequencing highlights the dynamic early history of dogs*](#). 2014. *PLoS Genet*. 10(1):e1004016. doi: 10.1371/journal.pgen.1004016.

2013

132. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W, Dutil J, Via M, Sandoval K, Bedoya G; 1000 Genomes Project, Oleksyk TK, Ruiz-Linares A, Burchard EG, Martinez-Cruzado JC, **Bustamante CD**. 2013. [*Reconstructing native american migrations from whole-genome and whole-exome data*](#). **PLoS Genet**. 9(12):e1004023. doi: 10.1371/journal.pgen.1004023.
131. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Grealley JM, Wang J, **Bustamante CD**, Boyko AR. [*Genetic recombination is targeted towards gene promoter regions in dogs*](#). 2013. **PLoS Genet**. 9(12):e1003984. doi: 10.1371/journal.pgen.1003984. PMID: 24348265.
130. Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li Y, Wang J, Gilbert MT, Willerslev E, Greenleaf WJ, **Bustamante CD**. [*Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries*](#). 2013. **Am J Hum Genet**. 7;93(5):852-64. doi: 10.1016/j.ajhg.2013.10.002. PMID: 24568772.
129. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, Eng C, Sandoval K, Acevedo-Acevedo S, Norman PJ, Layrisse Z, Parham P, Martínez-Cruzado JC, Burchard EG, Cuccaro ML, Martin ER, **Bustamante CD**. [*Reconstructing the population genetic history of the Caribbean*](#). 2013. **PLoS Genet**. 9(11):e1003925. doi: 10.1371/journal.pgen.1003925. Epub 2013 Nov 14. PMID: 24244192.
128. Goldstein O, Mezey JG, Schweitzer PA, Boyko AR, Gao C, **Bustamante CD**, Jordan JA, Aguirre GD, Acland GM. [*QCB1 and PDE6B mutations cause similar early onset retinal degenerations in two closely related terrier dog breeds*](#). 2013. *Invest Ophthalmol Vis Sci*. 54(10):7005-19. doi: 10.1167/iovs.13-12915. PMID: 24045995.
127. Kumar R, Nguyen EA, Roth LA, Oh SS, Gignoux CR, Huntsman S, Eng C, Moreno-Estrada A, Sandoval K, Peñaloza-Espinosa RI, López-López M, Avila PC, Farber HJ, Tcheurekdjian H, Rodriguez-Cintron W, Rodriguez-Santana JR, Serebrisky D, Thyne SM, Williams LK, Winkler C, **Bustamante CD**, Pérez-Stable EJ, Borrell LN, Burchard EG. [*Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics \(GALA II\) study*](#). 2013. **J Allergy Clin Immunol** 132(4):896-905.e1. doi: 10.1016/j.jaci.2013.02.046. PMID: 23684070.

126. Rootsi S, Behar DM, Järve M, Lin AA, Myres NM, Passarelli B, Poznik GD, Tzur S, Sahakyan H, Pathak AK, Rosset S, Metspalu M, Grugni V, Semino O, Metspalu E, **Bustamante CD**, Skorecki K, Vilems R, Kivisild T, Underhill PA. [Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites](#). 2013. **Nat Commun** 4:2928. doi: 10.1038/ncomms3928. PMID: 24346185.
125. Ma X, Kelley JL, Eilertson K, Musharoff S, Degenhardt JD, Martins AL, Vinar T, Kosiol C, Siepel A, Gutenkunst RN, **Bustamante CD**. 2013. [Population genomic analysis reveals a rich speciation and demographic history of orang-utans \(*Pongo pygmaeus* and *Pongo abelii*\)](#). **PLoS One** 23;8(10):e77175. doi: 10.1371/journal.pone.0077175. eCollection 2013. PMID: 24194868
124. Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. [RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference](#). **Am J Hum Genet.** 93(2):278-88 PMID: 23910464
123. Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, Snyder MP, Quintana-Murci L, Kidd JM, Underhill PA, Bustamante CD. 2013. [Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females](#). **Science** 341:562-5. PMID: 23908239
122. Prado-Martinez J, Sudmant PH, Kidd JM,..., **Bustamante CD**, Hammer MF, Eichler EE, Marques-Bonet T. 2013. [Great ape genetic diversity and population history](#). **Nature** 499:471—5.
121. Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, Bertranpetit J, Comas D, and **Bustamante CD**. 2013. [Gene flow from North Africa contributes to differential human genetic diversity in southern Europe](#). **PNAS** 110(29):11791—6. PMID: 23733930.
120. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D 3rd, Cornejo O, Findley SD, Zheng P, Utró F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, **Bustamante CD**, Schnell RJ, Main D, Gilbert D, Parida L, Kuhn DN. 2013. [The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color](#). **Genome Biol** 3;14(6):R53. PMID: 23731509.
119. Corona E, Chen R, Sikora M, Morgan AA, Patel CJ, Ramesh A, **Bustamante CD**, Butte AJ. 2013. [Analysis of the genetic basis of disease in the context of worldwide human relationships and migration](#). **PLoS Genet.** 9(5):e1003447. PMID: 23717210; PMCID: PMC3662561.
118. Kumar R, Nguyen EA, Roth LA, Oh SS, Gignoux CR, Huntsman S, Eng C, Moreno-Estrada A, Sandoval K, Peñaloza-Espinosa RI, López-López M, Avila PC, Farber HJ, Tcheurekdjian H, Rodriguez-Cintron W, Rodriguez-Santana JR, Serebrisky D, Thyne SM, Williams LK, Winkler C, **Bustamante CD**, Pérez-Stable EJ, Borrell LN, Burchard EG. 2013. [Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics \(GALA II\) study](#). **J Allergy Clin Immuno** S0091-6749(13)00586-1. PMID: 23684070.
117. Cornejo OE, Lefébure T, Bitar PD, Lang P, Richards VP, Eilertson K, Do T, Beighton D, Zeng L, Ahn SJ, Burne RA, Siepel A, **Bustamante CD**, Stanhope MJ. 2013. [Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*](#). **Mol Biol Evol** 30(4):881-93. PMID: 23228887.

116. Casto AM, Henn BM, Kidd JM, **Bustamante CD**, Feldman MW. 2012. [A tale of two haplotypes: the EDA2R/AR Intergenic region is the most divergent genomic segment between Africans and East Asians in the human genome.](#) **Hum Biol.** 84(6):641-94. PMID: 23959643.
115. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG, and **Bustamante CD**. 2012. [PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations.](#) **Hum Biol.** 84(4):343-64. PMID: 23249312.
114. Eilertson KE, Booth JG, and **Bustamante CD**. 2012. *SnIPRE: Selection Inference Using a Poisson Random Effects Model.* **PLoS Comput Biol** 8(12): e1002806. doi:10.1371/journal.pcbi.1002806.
113. 1000 Genomes Project Consortium. *An integrated map of genetic variation from 1,092 human genomes.* 2012. **Nature** 491:56-65. PMID: 23128226.
112. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh PR, Lachance J, Mountain J, **Bustamante CD**, Berger B, Tishkoff SA, Henn BM, Stoneking M, Reich D, and Pakendorf B. 2012. *The genetic prehistory of southern Africa.* **Nat Commun.** 3:1143. PMID: 23072811.
111. Sánchez-Quinto F, Botigué LR, Civit S, Arenas C, Avila-Arcos MC, **Bustamante CD**, Comas D, and Lalueza-Fox C. 2012. *North African populations carry the signature of admixture with Neandertals.* **PLoS One** 7(10):e47765. PMID: 23082212; PMCID: PMC3474783.
110. Kidd JM, Gravel G, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, McLaughlin SF, Peckham HE, Omberg L, Bormann Chung CA, Stanley S, Pearlstein K, Levandowsky E, Acevedo-Acevedo S, Auton A, Keinan A, Acuña-Alonzo V, Barquera-Lozano R, Canizales-Quinteros S, Eng S, Burchard EG, Russell A, Reynolds A, Clark AG, Reese MG, Lincoln SE, Butte AJ, De La Vega FM, and **Bustamante CD**. 2012. *Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation.* **Am J Hum Genet** 91(4):660-71. PMID: 23040495.
109. Granka JM, Henn BM, Gignoux CR, Kidd JM, **Bustamante CD**, and Feldman MW. 2012. *Limited Evidence for Classic Selective Sweeps in African Populations.* **Genetics.** PMID: 22960214 (cover article).
108. Kenny EE, Timpson, NJ, Sikora M., Yee, M-C, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, **Bustamante CD***, and Myles S.* 2012. *Melanesian blond hair is caused by an amino acid change in TYRP1.* **Science** 336(6081):554. PMID: 22556244.
107. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, **Bustamante CD**, Bamshad MJ, Akey JM; Broad GO; Seattle GO. 2012 *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.* **Science** 337(6090):64-9. PMID: 22604720.
106. Campbell CL, Palamara PF, Dubrovsky M, Botigué LR, Fellous M, Atzmon G, Oddoux C, Pearlman A, Hao L, Henn BM, Burns E, **Bustamante CD**, Comas D, Friedman E, Pe'er I, and Ostrer H. 2012. *North African Jewish and non-Jewish populations form distinctive, orthogonal clusters.* **Proc Natl Acad Sci U S A.** 2012 Aug 21;109(34):13865-70. PMCID:

PMC3427049.

105. Clark RT, Famoso AN, Zhao K, Shaff JE, Craft EJ, **Bustamante CD**, McCouch SR, Aneshansley DJ, and Kochian LV. 2012. *High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development*. **Plant Cell Environ.** (in press) PMID: 22860896.
104. Ma X, Rogacheva MV, Nishant KT, Zanders S, **Bustamante CD**, and Alani E. 2012. *Mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences*. **Cell Rep.** 1(1):36—42.
103. Kelley JL, Yee MC, Lee C, Levandowsky E, Shah M, Harkins T, Earley RL, and **Bustamante CD**. 2012. *The Possibility of De Novo Assembly of the Genome and Population Genomics of the Mangrove Rivulus *Kryptolebias marmoratus**. **Integr Comp Biol** 52(6):737-42 PMID: 22723055
102. Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R, Boyko AR, Fondon JW 3rd, Wayne RK, **Bustamante CD**, Ciruna B, and Ostrander EA. 2012. *Variation of BMP3 Contributes to Dog Breed Skull Diversity*. **PLoS Genet.** 8(8):e1002849. PMCID: PMC3410846.
101. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, **Bustamante CD**, and Snyder M. 2012. *Detecting and annotating genetic variations using the Hgaseq pipeline*. **Nat Biotechnol** 30(3):226—9. PMID: 22398614.
100. Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlouzi-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., and Comas, D. 2012. *Genomic ancestry of North Africans supports back-to-Africa migrations*. **PLoS Genet** 8(1): e1002397. PMCID: PMC3257290.
99. Chen, R., Corona, E., Sikora, M., Dudley, J. T., Morgan, A. A., Moreno-Estrada, A., Nilsen, G. B., Ruau, D., Lincoln, S. E., **Bustamante, C. D.**, and Butte, A. J. 2012. *Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases*. **PLoS Genet** 8(4):e1002621. PMCID: PMC3325177.
98. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, **Bustamante CD**, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A. 2012. *New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing*. **Nat Commun.** 3:698.

2011

97. Kenny EE, **Bustamante CD**. 2011. *SnapShot: Human biomedical genomics*. **Cell** 147(1):248- e1.
96. **Bustamante CD**, Burchard EG, De la Vega FM. 2011. *Genomics for the world*. **Nature** 475:163-5.
95. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK, Cornejo OE, Knowles JW, Woon M, Sangkuhl K, Gong L, Thorn CF, Hebert JM, Capriotti E, David SP, Pavlovic A, West A, Thakuria JV, Ball MP, Zaranek AW, Rehm HL, Church GM, West JS, **Bustamante CD**, Snyder M, Altman RB, Klein

- TE, Butte AJ, Ashley EA. 2011. *Phased whole-genome genetic risk in a family quartet using a major allele reference sequence*. **PLoS Genet** 7(9):e1002280. PMCID: 3174201.
94. Gao H, Bryc K, **Bustamante CD**. *On identifying the optimal number of population clusters via the deviance information criterion*. **PLoS One**. 2011;6(6):e21014. PMCID: 3125185.
 93. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, Chen Y, Challis D, Clarke L, Ball EV, Cibulskis K, Cooper DN, Fulton B, Hartl C, Koboldt D, Muzny D, Smith R, Sougnez C, Stewart C, Ward A, Yu J, Xue Y, Altshuler D, **Bustamante CD**, Clark AG, Daly M, Depristo M, Flicek P, Gabriel S, Mardis E, Palotie A, Gibbs RA, 1000 Genomes Project. *The functional spectrum of low-frequency coding variation*. **Genome Biol**. 2011;12(9):R84.
 92. Rasmussen M, Guo X, Wang Y, Lohmueller KE,..., **Bustamante CD**,..., Nielsen R, Jun W, Willerslev E. *An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia*. **Science**. 2011.
 91. Xie X, Molina J, Hernandez R, Reynolds A, Boyko AR, **Bustamante CD**, Purugganan MD. *Levels and patterns of nucleotide variation in domestication QTL regions on rice chromosome 3 suggest lineage-specific selection*. **PLoS One**. 2011;6(6):e20670. PMCID: 3108957.
 90. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, **Bustamante CD***, McCouch SR*. *Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa**. **Nat Commun**. 2011;2:467.
 89. Vonholdt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jedrzejewski W, Jedrzejewska B, Sidorovich V, Greco C, Randi E, Musiani M, Kays R, **Bustamante CD**, Ostrander EA, Novembre J, Wayne RK. *A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids*. **Genome Res**. 2011.
 88. Brisbin A, Cruickshank J, Moise NS, Gunn T, **Bustamante CD**, Mezey JG. *Fast, exact linkage analysis for categorical traits on arbitrary pedigree designs*. **Genet Epidemiol**. 2011.
 87. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, Huang P, Jackson S, Schaal BA, **Bustamante CD**, Boyko AR, Purugganan MD. *Molecular evidence for a single evolutionary origin of domesticated rice*. **Proc Natl Acad Sci U S A**. 2011.108(20):8351-6.
 86. Myers RA, Casals F, Gauthier J, Hamdan FF, Keebler J, Boyko AR, **Bustamante CD**, Piton AM, Spiegelman D, Henrion E, Zilvermit M, Hussin J, Quinlan J, Yang Y, Lafreniere RG, Griffing AR, Stone EA, Rouleau GA, Awadalla P. *A population genetic approach to mapping neurological disorder genes using deep resequencing*. **PLoS Genet**. 2011;7(2):e1001318. PMCID: 3044677.
 85. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, **Bustamante CD**. *Demographic history and rare allele sharing among human populations*. **Proc Natl Acad Sci U S A**. 2011;108(29):11983-8. PMCID: 3142009.
 84. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A, Lin AA, Underhill PA, Comas D, Kidd KK, Norman PJ, Parham P, **Bustamante CD**, Mountain JL, Feldman MW. *Feature Article: Hunter-gatherer genomic diversity suggests a southern African origin for modern humans*. **Proc Natl Acad Sci U S A**. 2011. 108(13):5154-62. PMCID: 3069156.
 83. Locke DP, Hillier LW, Warren WC, ... , **Bustamante CD**, ... , Gibbs RA, Wilson RK.

Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469(7331):529-33. PMID: 3060778.

82. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, **Bustamante CD***, Buckler ES*. *From the Cover: Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A.* 2011;108(9):3530-5. PMID: 3048109.
81. Lohmueller KE, **Bustamante CD**, Clark AG. *Detecting directional selection in the presence of recent admixture in African Americans. Genetics.* 2011;187(3):823-35.

2010

77. Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin SQ, Hurle B, Schwartzberg PL, Williamson SH, **Bustamante CD**, Nielsen R, Clark AG, Green ED. *Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet.* 2010;6(10):e1001157. PMID: 2954825.
76. Wright MH, Tung CW, Zhao K, Reynolds A, McCouch SR, **Bustamante CD**. *ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. Bioinformatics.* 2010;26(23):2952-60. PMID: 2982150.
75. Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, **Bustamante CD**. *Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet.* 2010;19(R2):R221-6. PMID: 2953744.
74. Brisbin A, Weissman MM, Fyer AJ, Hamilton SP, Knowles JA, **Bustamante CD**, Mezey JG. *Bayesian linkage analysis of categorical traits for arbitrary pedigree designs. PLoS One.* 2010;5(8):e12307. PMID: 2928726.
73. Nishant KT, Wei W, Mancera E, Argueso JL, Schlattl A, Delhomme N, Ma X, **Bustamante CD**, Korbel JO, Gu Z, Steinmetz LM, Alani E. *The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. PLoS Genet.* 2010;6(9). PMID: 2936533.
72. Goldstein O, Mezey JG, Boyko AR, Gao C, Wang W, **Bustamante CD**, Anguish LJ, Jordan JA, Pearce-Kelling SE, Aguirre GD, Acland GM. *An ADAM9 mutation in canine cone-rod dystrophy 3 establishes homology with human cone-rod dystrophy 9. Mol Vis.* 2010;16:1549-69. PMID: 2925905.
71. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, Cargill M, Auton A, Reynolds A, Elkahoul AG, Castelhamo M, Mosher DS, Sutter NB, Johnson GS, Novembre J, Hubisz MJ, Siepel A, Wayne RK, **Bustamante CD**, Ostrander EA. *A simple genetic architecture underlies morphological variation in dogs. PLoS Biol.* 2010;8(8):e1000451. PMID: 2919785.
70. Zanders S, Ma X, Roychoudhury A, Hernandez RD, Demogines A, Barker B, Gu Z, **Bustamante CD**, Alani E. *Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a Bayesian approach. Genetics.* 2010;186(2):493-503. PMID: 2954485.
69. Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, **Bustamante CD**, McCouch SR. *Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. PLoS One.* 2010;5(5):e10780. PMID: 2875394.

68. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, **Bustamante CD**, Ostrer H. *Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations*. **Proc Natl Acad Sci U S A**. 2010;107 Suppl 2:8954-61. PMCID: 3024022.
67. Williams LM, Ma X, Boyko AR, **Bustamante CD**, Oleksiak MF. *SNP identification, verification, and utility for population genetics in a non-model genus*. **BMC Genet**. 2010;11:32. PMCID: 2874759.
66. Brideau CM, Eilertson KE, Hagarman JA, **Bustamante CD**, Soloway PD. *Successful computational prediction of novel imprinted genes from epigenomic features*. **Mol Cell Biol**. 2010;30(13):3357-70. PMCID: 2897571.
65. Lohmueller KE, **Bustamante CD**, Clark AG. *The effect of recent admixture on inference of ancient human population history*. **Genetics**. 2010;185(2):611-22. PMCID: 2881141.
64. Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, Reynolds A, Bryc K, Brisbin A, Knowles JC, Mosher DS, Spady TC, Elkahouloun A, Geffen E, Pilot M, Jedrzejewski W, Greco C, Randi E, Bannasch D, Wilton A, Shearman J, Musiani M, Cargill M, Jones PG, Qian Z, Huang W, Ding ZL, Zhang YP, **Bustamante CD**, Ostrander EA, Novembre J, Wayne RK. *Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication*. **Nature**. 2010;464(7290):898-902.
63. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, **Bustamante CD**. *Genome-wide patterns of population structure and admixture in West Africans and African Americans*. **Proc Natl Acad Sci U S A**. 2010;107(2):786-91. PMCID: 2818934.

2009

62. Gutenkunst RN, Hernandez RD, Williamson SH, **Bustamante CD**. *Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data*. **PLoS Genet**. 2009;5(10):e1000695. PMCID: 2760211.
61. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, Vonholdt BM, Rhue A, Boyko A, Byers A, Wong A, Mosher DS, Elkahouloun AG, Spady TC, Andre C, Lark KG, Cargill M, **Bustamante CD**, Wayne RK, Ostrander EA. *Coat variation in the domestic dog is governed by variants in three genes*. **Science**. 2009;326(5949):150-3. PMCID: 2897713.
60. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, **Bustamante CD**, Clark AG, Nielsen R. *Targets of balancing selection in the human genome*. **Mol Biol Evol**. 2009;26(12):2755-64. PMCID: 2782326.
59. Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhana M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R, Ostrander EA, Schoenebeck J, Todhunter RJ, Jones P, **Bustamante CD**. *Complex population structure in African village dogs and its implications for inferring dog domestication history*. **Proc Natl Acad Sci U S A**. 2009;106(33):13903-8. PMCID: 2728993.
58. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, **Bustamante CD**, Clark AG. *Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence*. **PLoS Genet**. 2009;5(8):e1000592. PMCID: 2714078.

57. Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahouloun A, Cargill M, Jones PG, Maslen CL, Acland GM, Sutter NB, Kuroki K, **Bustamante CD**, Wayne RK, Ostrander EA. *An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs.* **Science**. 2009;325(5943):995-8. PMID: 2748762.
56. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, **Bustamante CD**, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell CR, Leung H, Leach JE. *Genomewide SNP variation reveals relationships among landraces and modern varieties of rice.* **Proc Natl Acad Sci U S A**. 2009;106(30):12273-8. PMID: 2718348.
55. Gibbs RA, Taylor JF, Van Tassell CP, ... , **Bustamante CD**, ..., Dodds KG. *Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds.* **Science**. 2009;324(5926):528-32. PMID: 2735092.
54. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, **Bustamante CD**, Clark AG. *Darwinian and demographic forces affecting human protein coding genes.* **Genome Res**. 2009;19(5):838-49. PMID: 2675972.
53. Santos VJ, **Bustamante CD**, Valero-Cuevas FJ. *Improving the fitness of high-dimensional biomechanical models via data-driven stochastic exploration.* **IEEE Trans Biomed Eng**. 2009;56(3):552-64. PMID: 2841988.
52. Lohmueller KE, **Bustamante CD**, Clark AG. *Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data.* **Genetics**. 2009;182(1):217-31. PMID: 2674818.
51. Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, **Bustamante CD**. *Global distribution of genomic diversity underscores rich complex history of continental human populations.* **Genome Res**. 2009;19(5):795-803. PMID: 2675968.
50. Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, **Bustamante CD**, Ostrander EA, Tang H, Wayne RK, Barsh GS. *Molecular and evolutionary history of melanism in North American gray wolves.* **Science**. 2009;323(5919):1339-43. PMID: 2903542.
49. Gray MM, Granka JM, **Bustamante CD**, Sutter NB, Boyko AR, Zhu L, Ostrander EA, Wayne RK. *Linkage disequilibrium and demographic history of wild and domestic canids.* **Genetics**. 2009;181(4):1493-505. PMID: 2666515.
48. Degenhardt JD, de Candia P, Chabot A, Schwartz S, Henderson L, Ling B, Hunter M, Jiang Z, Palermo RE, Katze M, Eichler EE, Ventura M, Rogers J, Marx P, Gilad Y, **Bustamante CD**. *Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in Rhesus Macaques (Macaca mulatta).* **PLoS Genet**. 2009;5(1):e1000346. PMID: 2621346.
47. **Bustamante CD**, Ramachandran S. *Evaluating signatures of sex-specific processes in the human genome.* **Nat Genet**. 2009;41(1):8-10. PMID: 2766813.

2008

46. Brooks MB, Catalfamo JL, Etter K, Brisbin A, **Bustamante CD**. *Exclusion of ABCA-1 as a*

candidate gene for canine Scott syndrome. J Thromb Haemost. 2008;6(9):1608-10. PMID: 2766810.

45. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, **Bustamante CD**. *Genes mirror geography within Europe. Nature.* 2008;456(7218):98-101. PMID: 2735096.
44. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, **Bustamante CD**, Roses AD, Burns DK, Ehm MG, Lai EH. *The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet.* 2008;83(3):347-58. PMID: 2556436.
43. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, **Bustamante CD**, Nielsen R, Siepel A. *Patterns of positive selection in six Mammalian genomes. PLoS Genet.* 2008;4(8):e1000144. PMID: 2483296.
42. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, **Bustamante CD**, Teshima KM, Przeworski M. *Natural selection on genes that underlie human disease susceptibility. Curr Biol.* 2008;18(12):883-9. PMID: 2474766.
41. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, **Bustamante CD**. *Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet.* 2008;4(5):e1000083. PMID: 2377339.
40. Padhukasahasram B, Marjoram P, Wall JD, **Bustamante CD**, Nordborg M. *Exploring population genetic models with recombination using efficient forward-time simulations. Genetics.* 2008;178(4):2417-27. PMID: 2323826.
39. Shimizu KK, Reininga JM, Caicedo AL, Mays CA, Moore RC, Olsen KM, Ruzsa S, Coop G, **Bustamante CD**, Purugganan MD. *Retraction. Science.* 2008;320(5873):176.
38. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, **Bustamante CD**. *Proportionally more deleterious genetic variation in European than in African populations. Nature.* 2008;451(7181):994-7. PMID: 2923434.
37. Huerta-Sanchez E, Durrett R, **Bustamante CD**. *Population genetics of polymorphism and divergence under fluctuating selection. Genetics.* 2008;178(1):325-37. PMID: 2206081.

2007

36. Nielsen R, Hellmann I, Hubisz M, **Bustamante C**, Clark AG. *Recent and ongoing selection in the human genome. Nat Rev Genet.* 2007;8(11):857-68. PMID: 2933187.
35. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fedel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, **Bustamante CD**, Purugganan MD. *Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet.* 2007;3(9):1745-56. PMID: 1994709.
34. Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, **Bustamante CD**, McCouch SR. *Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet.*

2007;3(8):e133. PMCID: 1941752.

33. Hernandez RD, Williamson SH, Zhu L, **Bustamante CD**. *Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans*. **Mol Biol Evol**. 2007;24(10):2196-202.
32. Jensen JD, Thornton KR, **Bustamante CD**, Aquadro CF. *On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations*. **Genetics**. 2007;176(4):2371-9. PMCID:1950638.
31. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, **Bustamante CD**, Nielsen R. *Localizing recent adaptive evolution in the human genome*. **PLoS Genet**. 2007;3(6):e90. PMCID: 1885279.
30. Hernandez RD, Williamson SH, **Bustamante CD**. *Context dependence, ancestral misidentification, and spurious signatures of natural selection*. **Mol Biol Evol**. 2007;24(8):1792-800.
29. Mosher DS, Quignon P, **Bustamante CD**, Sutter NB, Mellersh CS, Parker HG, Ostrander EA. *A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs*. **PLoS Genet**. 2007;3(5):e79. PMCID: 1877876.
28. Gao H, Williamson S, **Bustamante CD**. *A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data*. **Genetics**. 2007;176(3):1635-51. PMCID: 1931536.
27. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, Nazareth L, Indap A, Bourquin T, McPherson J, Muzny D, Gibbs R, Nielsen R, **Bustamante CD**. *Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques*. **Science**. 2007;316(5822):240-3.
26. Gibbs RA, Rogers J, Katze MG, ... , **Bustamante CD**, ..., Smith KE, Zieg AS. *Evolutionary and biomedical insights from the rhesus macaque genome*. **Science**. 2007;316(5822):222-34.
25. Sutter NB, **Bustamante CD**, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG, Quignon P, Johnson GS, Parker HG, Fretwell N, Mosher DS, Lawler DF, Satyaraj E, Nordborg M, Lark KG, Wayne RK, Ostrander EA. *A single IGF1 allele is a major determinant of small size in dogs*. **Science**. 2007;316(5821):112-5. PMCID: 2789551.

Pre-2006

23. Pollinger JP, **Bustamante CD**, Fledel-Alon A, Schmutz S, Gray MM, Wayne RK. *Selective sweep mapping of genes with large phenotypic effects*. **Genome Res**. 2005;15(12):1809-19. PMCID: 1356119.
22. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, **Bustamante C**. *Genomic scans for selective sweeps using SNP data*. **Genome Res**. 2005;15(11):1566-75. PMCID: 1310644.
21. Clark AG, Hubisz MJ, **Bustamante CD**, Williamson SH, Nielsen R. *Ascertainment bias in studies of human genome-wide polymorphism*. **Genome Res**. 2005;15(11):1496-502. PMCID: 1310637.
20. **Bustamante CD**, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M,

- Clark AG. *Natural selection on protein-coding genes in the human genome*. **Nature**. 2005;437(7062):1153-7.
19. Dimmic MW, Hubisz MJ, **Bustamante CD**, Nielsen R. *Detecting coevolving amino acid sites using Bayesian mutational mapping*. **Bioinformatics**. 2005;21 Suppl 1:i126-35.
 18. Jensen JD, Kim Y, DuMont VB, Aquadro CF, **Bustamante CD**. *Distinguishing between selective sweeps and demography using DNA polymorphism data*. **Genetics**. 2005;170(3):1401-10. PMID: 1451184.
 17. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, **Bustamante CD**. *Simultaneous inference of selection and population growth from patterns of variation in the human genome*. **Proc Natl Acad Sci U S A**. 2005;102(22):7882-7. PMID: 1142382.
 16. Zhu L, **Bustamante CD**. *A composite-likelihood approach for detecting directional selection from DNA sequence data*. **Genetics**. 2005;170(3):1411-21. PMID: 1451173.
 15. Nielsen R, **Bustamante C**, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J JS, Adams MD, Cargill M. *A scan for positively selected genes in the genomes of humans and chimpanzees*. **PLoS Biol**. 2005;3(6):e170. PMID: 1088278.
 14. Dimmic MW, Sunyaev S, **Bustamante CD**. *Inferring SNP function using evolutionary, structural, and computational methods*. **Pac Symp Biocomput**. 2005:382-4.
 13. Shimizu KK, Cork JM, Caicedo AL, Mays CA, Moore RC, Olsen KM, Ruzsa S, Coop G, **Bustamante CD**, Awadalla P, Purugganan MD. *Darwinian selection on a selfing locus*. **Science**. 2004;306(5704):2081-4.
 12. Williamson S, Perry SM, **Bustamante CD**, Orive ME, Stearns MN, Kelly JK. *A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients*. **Mol Biol Evol**. 2005;22(3):456-68.
 11. Williamson S, Fledel-Alon A, **Bustamante CD**. *Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance*. **Genetics**. 2004;168(1):463-75. PMID: 1448126.
 10. Sawyer SA, Kulathinal RJ, **Bustamante CD**, Hartl DL. *Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection*. **J Mol Evol**. 2003;57 Suppl 1:S154-64.
 9. Gilad Y, **Bustamante CD**, Lancet D, Paabo S. *Natural selection on the olfactory receptor gene family in humans and chimpanzees*. **Am J Hum Genet**. 2003;73(3):489-501. PMID: 1180675.
 8. Barrier M, **Bustamante CD**, Yu J, Purugganan MD. *Selection on rapidly evolving proteins in the Arabidopsis genome*. **Genetics**. 2003;163(2):723-33. PMID: 1462452.
 7. **Bustamante CD**, Nielsen R, Hartl DL. *Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data*. **Theor Popul Biol**. 2003;63(2):91-103.
 6. **Bustamante CD**, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. *The cost of inbreeding in Arabidopsis*. **Nature**. 2002;416(6880):531-4.
 5. **Bustamante CD**, Wakeley J, Sawyer S, Hartl DL. *Directional selection and the site-frequency spectrum*. **Genetics**. 2001;159(4):1779-88. PMID: 1461920.

4. **Bustamante CD**, Nielsen R, Hartl DL. *A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents.* **Mol Biol Evol.** 2002;19(1):110-7.
3. Nurminsky D, Aguiar DD, **Bustamante CD**, Hartl DL. *Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*.* **Science.** 2001;291(5501):128-30.
2. **Bustamante CD**, Townsend JP, Hartl DL. *Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*.* **Mol Biol Evol.** 2000;17(2):301-8.
1. Jimenez JJ, Alvarez E, **Bustamante CD**, Yunis AA. *Pretreatment with 1,25(OH)2D3 protects from Cytoxan-induced alopecia without protecting the leukemic cells from Cytoxan.* **Am J Med Sci.** 1995;310(2):43-7.

Commentaries and non-peer reviewed publications (6)

248. Barnes KC, De La Vega FM, **Bustamante CD**, Gignoux CR, Kenny E, Mathias RA, Pasaniuc B. [Session Introduction: Overcoming health disparities in precision medicine.](#) **Pac Symp Biocomput.** 2023;28:181-185. PMID: 36540975.
216. De La Vega FM, **Bustamante CD**. [Polygenic risk scores: a biased prediction?](#) **Genome Med.** 2018 Dec 27;10(1):100. doi: 10.1186/s13073-018-0610-x. PMID: 30591078
155. **Bustamante CD**, Rasmussen M. 2015. [Beyond the reference genome.](#) **Nat Biotechnol.** 2015 Jun;33(6):605-6. doi: 10.1038/nbt.3249. No abstract available. PMID: 26057977.
79. **Bustamante CD**, Henn BM. *Human origins: Shadows of early migrations.* 2010. **Nature** 468:1044-5.
78. De La Vega FM, **Bustamante CD**, Leal SM. *Genome-Wide Association Mapping And Rare Alleles: From Population Genomics To Personalized Medicine - Session Introduction.* **Pac Symp Biocomput.** 2011:74-5.
24. Abecasis G, Tam PK, **Bustamante CD**, Ostrander EA, Scherer SW, Chanock SJ, Kwok PY, Brookes AJ. *Human Genome Variation 2006: emerging views on structural variation and large-scale SNP analysis.* **Nat Genet.** 2007;39(2):153-5.

Book Chapters

1. Kulathinal R, Sawyer SA, **Bustamante CD**, Nurminsky DI, Ponce R, Ranz JM, Hartl DL. *Selective Sweep in the Evolution of a New Sperm-Specific Gene in *Drosophila*.* In: Nurminsky D, editor. *Selective Sweeps*; Springer; 2005.
2. Przeworski M, **Bustamante CD**. *Genetic signatures of Natural Selection.* *Encyclopedia of Genetics.* London, U.K.: Wiley & Sons; 2005.
3. Ramachandran S, Tang H, Gutenkunst R, **Bustamante C**. *Genetics and Genomics of Human Population Structure.* In: Speicher MR, Motulsky AG, Antonarakis SE, editors. *Vogel and Motulsky's Human Genetics: Problems and Approaches.* Hiedelberg: Springer; 2010. p. 589-615.

Research Support (Active)

None

Research Support (Pending)

Galatea Bio, Inc. has multiple SBIR grants under consideration with NIH

Research Support (Completed)

U41HG009649 (Bustamante [C])

08/01/2017 – 07/31/2021

SPO# 109590

NIH

Clinical Genome Resource (ClinGen)

Major goal is to create a unified, public, and freely available database of genetic alterations relevant to clinical care. Our ultimate goal is to empower clinicians, genetic counselors, and patients to make informed decisions based on DNA testing.

Role: PI

1U01HG009080 (Bustamante [C])

05/20/2016 – 03/31/2020

SPO# 121626

NIH

Center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Disease

Major goal is to accelerate the discovery of DNA variation relevant to health and disease by analyzing data from over 225,000 ethnically and racially diverse patients that will undergo genome sequencing. Of particular importance is ensuring we have powerful statistical methods for analyzing data from underserved groups including U.S. minority populations. Achieving this goal requires expertise across many domains of knowledge including: medical and population genomics, algorithm development for disease mapping, and expertise in large-scale databases.

Role: PI

5R24 HL117756-02 (Wu)

04/15/2014 – 03/31/2019

NIH/NHLBI

SPO# 110465/ Cardiovascular Institute Operations

Biorepository of Human iPSCs for Studying Dilated and Hypertrophic Cardiomyopathy

The project seeks to create a bank of 600 iPSC lines created from cardiovascular patients and distribute to the NHLBI community.

Role: Co-PI

1U54MD01724-01 (Cullen[C], Bustamante)

04/01/2016 – 03/31/2021

SPO# 121927

NIH/NIA

Stanford Precision Health for Ethnic and Racial Equity (SPHERE)

Major goal is to provide services to the different projects in the center as well as any other group in the consortium. We already have a robust infrastructure including equipment, expertise, and personnel for performing state-of-the-art genotyping, genome and exome sequencing, transcriptomics, metabolomics, and microbiome characterization. These activities will be implemented into the individual projects.

Role: Lab Core PI

1U24EB02367401 (Ashley) 12/12/2016 – 11/30/2022
SPO# 124803
NIH
Stanford MoTrPAC Bioinformatics Center
Major goal is to manage the bioinformatics for the MoTrPAC Clinical Site
Role: Co-PI

1R01HG008150-01 (Montgomery) 08/01/2015 – 06/30/2018
SPO# 114738
NIH
SPO# 114738/ Pathology Research
Predicting casual non-coding variants in a founder population
The project seeks to develop methods for interpreting non-coding genetic variation: to provide a Bayesian method for predicting causal variants from full genomes, to identify and validate a large set of functional non-coding variants using CRISPR technology, and to predict disease-relevant traits likely to be affected by each variant.
Role: Co-PI

1U24HG008956-01 (Matis [C], Bustamante) 01/14/2016 – 11/30/2019
SPO# 120615
NIH
Rutgers University
NHGRI Genome Sequencing Program Coordinating Center
Major goal is to identify of relevant control sets for data sets coming through GSPCC, and on the use of diverse trans-ethnic populations in boosting power, identifying novel causal variants, and characterizing the genetic background of the common control resource.
Role: Subcontract PI

IOS-1642184 (Bustamante) 08/01/2016 – 07/31/2018
SPO# 124489
NSF
EAGER: Establishing the genetic basis of hibernation by building and utilizing a next-generation genomics resource for the model hibernator, the thirteen-lined ground squirrel
Major goal is to establish a genetic basis for the hibernating phenotype in the 13-lined ground squirrel by developing a high-quality genome resource for the species. Specifically, we will employ quantitative trait loci analysis to identify genetic variants associated with hibernation to determine whether hibernation-related traits are heritable.
Role: PI

U01FD004979 (Altman) 02/01/2015 – 03/31/2018
SPO# 119541
UCSF
UCSF-Stanford CERSI: Improving the Efficiency and Rigor of Pharmacovigilance (PV) at FDA
Role: Co-PI

4U01HG0074919-04 (Matis [C], Bustamante) 09/01/2013 – 05/31/2018
SPO# 109405
Rutgers University (Prime NIH/NHGRI)

NHGRI PAGE Coordinating Center

The major goals of this study are to provide flexible support in data analysis, management, synthesis, and dissemination, and program logistics, for the PAGE II studies.

Role: Subcontract PI

4U01HG007417 (Loos [C], Bustamante)

09/01/2013 – 05/31/2018

SPO# 109391

MSSM (Prime NIH/NHGRI)

Diverse ancestry biobank to map biomedical traits and elucidate health disparities

The major goal of this project is to characterize the population architecture of differences in disease burden among populations in a multi-ethnic cohort using common and rare coding variants. We will also develop methods for disease mapping based on local ancestry information and will assess aggregate effects of rare variants in the context of the PAGE II network.

Role: Co-PI

66864T (Bustamante, Martin, Pls)

09/01/2011 – 08/31/2017

NIH/NIGMS

Genomic Origins and Admixture in Latinos (GOAL)

The major goal of this project is to generate and analyze dense genotype data from 180 parent-offspring triads of diverse Hispanic/Latino ancestry sampled in Miami, as well as sequence the complete genomes of 30 H/L triads.

ED-9697-STS2 (Bustamante, Dermitzakis[C], Guigo, McCarthy) 08/01/2013 – 06/30/2016

NIH

Methods for High-Resolution Analysis of Genetic Effects on Gene Expression

The major goal of this project is to identify genetic effects on gene expression using GTEx data. Specifically, we will provide statistical expression quantitative trait methodology, network methods for detecting gene regulatory interactions, and bioinformatics methods for improving identification of allele-specific expression.

0255-6424-4609 (Loos, Bustamante)

09/01/2013 - 05/31/2017

NIH/NHGRI

Diverse Ancestry Biobank to Map Biomedical Traits and Elucidate Health Disparities

The major goal of this project is to characterize the population architecture of differences in disease burden among populations in a multi-ethnic cohort using common and rare coding variants. We will also develop methods for disease mapping based on local ancestry information and will assess aggregate effects of rare variants in the context of the PAGE II network.

5943 (Matisse)

09/01/2013 - 05/31/2017

NIH/NHGRI

NHGRI PAGE Coordinating Center

The major goals of this study are to provide flexible support in data analysis, management, synthesis, and dissemination, and program logistics, for the PAGE II studies.

2R01HG003229 (Clark, Bustamante, Nielsen, Pls)

01/01/2010 – 6/30/2013

NIH/NHGRI

Population Genetic Inferences from Dense Genotype Data

The major goals of this project are to develop methods for population genetic inference from next-generation sequence data, including ascertainment bias correction, inference of natural selection, and prediction of functionality for novel SNPs and CNVs in the human genome.

U01HG005715 (Bustamante, PI)

09/09/2010 – 06/30/2014

NIH/NHGRI

Population Structure Admixture and Selection across the 1000 Genomes Data Set

The main goals of the project are to develop novel approaches for characterizing population structure, analyzing patterns of admixture, and localizing signatures of selection across the 2,500 samples sequenced by the 1000 Genomes Project.

58-6631-0-100 (Bustamante, PI)

07/01/2010 – 01/07/2014

USDA

Identification of Genes Regulating Major Agronomic Traits in Cacao Through Association and Linkage Mapping Using Large Sequence Data Sets

The major goal of this project is provide analytical and bioinformatic support to the MARS-USDA *Theobroma cacao* L. sequencing project, including SNP diversity and association mapping analysis of cacao populations as well as gene expression analysis of diverse *Theobroma* species.

1R01GM083606 (Bustamante, PI)

08/01/2007–07/31/2012

NIH/NIGMS

Association Mapping in Structured Populations

The major goal of this project is to develop novel statistical methods for association mapping in structured populations and to use these methods to map several complex traits in ethnically diverse African populations.

1R01MH084695 (Clark, Bustamante, Nielsen, PIs)

09/28/2008–07/31/2012

NIH/NIMH

Inferring multiple-SNP disease association with DNA resequence data

The major goal of this project is to develop novel statistical methods for association mapping of genetic variants underlying risk for complex disease, using resequenced DNA samples from cases and controls.

DBI 0701382 (Purugganan, PI, Bustamante, Co-PI)

09/01/2007–08/31/2012

NSF/DBI

GEPR: The Evolutionary Genomics of Rice Domestication

The major goals of this project are to fine-map domestication genes and examine the role of evolutionary selection in the origin of cultivated rice.

DEB 0516310 (Bustamante)

09/01/2005 – 08/31/2010

NSF

Computational Methods for Detecting Natural Selection using Comparative Population Genomic Data

The major goal of this project is to develop novel population genetic methods for identifying genes and genomic regions that are involved in adaptive molecular evolution through the comparison both within and between species genomic variation data.

5U01HL084706 -03 (Clark, PI, Bustamante, Co-PI)

06/15/2006 – 05/31/2010

NIH/NHLBI

Dimension Reduction Approaches for Genome-Wide Association Testing

The major goal of this project is to develop novel Bayesian shrinkage methods for genome-wide association mapping and to apply these methods to genome-wide expression level and genotype data from the 270 HapMap cell lines in order to map expression quantitative trait loci (eQTLs).

BIO 0948510 (Bustamante, PI)
NSF/DBI/EAGER

08/15/2009–7/31/2011 (NCE)

Genome-wide diversity in village dogs: uncovering the history of dog domestication, the genetics of natural dog populations, and consequences of artificial selection in breed dogs.

The major goals of this project are to develop population genetic methods for inferring the evolutionary history of dog domestication from genotyping and sequencing data, and to determine the impact of dog domestication on the genetic basis of complex traits in dogs.

DBI 0606461 (McCouch, PI, Bustamante, Co-PI) 08/01/2006 – 07/31/2011 (NCE) NSF/DBI
GEPR: Exploring the Genetic Basis of Transgressive Variation in Rice

The major goal of this project is to perform association mapping in rice (*Oryza sativa*) using 500 individuals genotyped at 10,000 SNPs in order to identify quantitative trait loci underlying 26 agronomic traits of interests.

1R01AI073368 (Stanhope, PI, Bustamante, Co-PI) 5/1/2010–6/15/2011 NIH/NIAID
Evolutionary Genomics and Population Genetics of Pathogenic Streptococci

The major goal of this project is to reach a thorough understanding of the molecular specifics correlated with adaptive differences within and between the pathogenic taxa of the genus *Streptococcus*.

673446Z (Nickerson, PI, Bustamante, Co-PI) 09/30/2009 - 06/30/2012 NIH
Northwest Genomic Center

The main goals of my subcontract on this project is to develop population genetic methods for quantifying population structure from whole-genome exome data and estimate selection coefficients on different categories of functional sites in the human genome.

Educational and Outreach Activities

Post-doctoral fellows trained

1. Dr. Adam Auton, D. Phil. (10/2007—6/2009), Statistical Geneticist, 23andme
2. Dr. Maria Avila-Arcos, Ph.D. (05/2014—12/2017), Assistant Professor, National Autonomous University of Mexico
3. Dr. Linda Boettger (Goodman), Ph.D. (1/2018 – 8//2018), Co-founder and CTO at Fauna Bio
4. Dr. Adam Boyko, Ph. D. (9/2005—8/2011), Associate Professor, Cornell University
5. Dr. Jake Byrnes, Ph. D. (9/2010—10/2011), Population Genomics Manager, Ancestry.com
6. Dr. Meredith Carpenter, Ph. D. (7/2011—1/2014), Chief Scientific Officer, ArcBio LLC
7. Dr. Jessica Chen, Ph.D. (3/2017 – 9/2019), Research Scientist, Genentech
8. Dr. Omar Cornejo, Ph. D. (8/2009—8/2013), Assistant Professor, Washington State University
9. Dr. Christopher DeBoever, Ph.D. (10/2016—4/2018), Vice President of Platform at Data4Cure, Inc.
10. Dr. Matthew J. Dimmic, Ph. D. (3/2003—2005), Genomics and Data Science Lead, Monsanto
11. Dr. Anna Katharina Elmer, Ph.D. (4/2014—9/2014), Senior Scientist, ArcBio LLC
12. Dr. Rosa Lorenzo Fregel, Ph.D. (6/2014—8/2017), Assistant Professor Tenerife
13. Dr. Christopher Gignoux, Ph.D. (8/2013—2/2017), Associate Professor, University of Colorado
14. Dr. Ryan Gutenkunst, Ph. D. (1/2008—12/2008), Assistant Professor, University of Arizona
15. Dr. Katherine Grabek, Ph.D. (7/2014—8/2018), Co-founder and CSO at Fauna Bio
16. Dr. Simon Gravel, Ph.D. (10/2009—12/2012), Assistant Professor, McGill University
17. Dr. Ragan Hart, Ph.D. (9/2018—6/2019), Entrepreneur-In-Residence, F-Prime Capital
18. Dr. Brenna Henn, Ph. D. (12/2009—12/2012), Assistant Professor, SUNY Stony Brook
19. Dr. Patrick House, Ph. D. (9/2014—9/2015), Neuroscientist and writer

20. Dr. Aashish Jha, Ph.D. (9/2015—6/2019), Director of Bioinformatics, NomNomNow, Berkeley, CA
21. Dr. Alex Ioannidis, Ph.D. (9/2018—6/2022), Postdoctoral scholar, Stanford
22. Dr. Joanna Kelley, Ph. D. (8/2010—8/2013), Assistant Professor, Washington State University
23. Dr. Eimear Kenny, Ph. D. (10/2010—10/2012), Professor, Mount Sinai School of Medicine
24. Dr. Jeffrey Kidd, Ph. D. (2/2010—12/2011), Associate Professor, University of Michigan School of Medicine
25. Dr. Carolin Kosiol, Ph. D. (9/2006—1/2009), Bioinformatics Group Leader, University of Veterinary Medicine (vetmeduni) Vienna, Austria
26. Dr. Tuuli Lappalainen (1/2013—12/2013), Professor, Department of Systems Biology, Columbia University and Junior Investigator and Core Member, New York Genome Center
27. Dr. Arturo Lopez Pineda (1/2016—4/2019), CEO, Amphora Health, SA
28. Dr. Fernando Mendez, Ph.D., (7/2013 – 7/2017), Senior Scientist - Population Geneticist at Helix
29. Dr. Babak Moatamed Ph.D. (7/2020 – 3/2022), Director of Clinical Informatics, Galatea Bio, Inc.
30. Dr. Nilah Monnier Ioannidis, Ph.D. (9/2013—10/2019), Assistant Professor, U. California Berkeley
31. Dr. Andres Moreno Estrada, M.D./Ph. D. (3/2009—12/2014), Professor and Group Leader, Langebio (Mexico)
32. Dr. Sonia Moreno Grau, Ph.D. (10/2020-1/2022), Director of Collaborations, Galatea Bio, Inc.
33. Dr. Marina Muzzio, Ph. D. (8/2011—8/2013), Researcher, Instituto Multidisciplinario de Biología Celular (IMBICE), National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina
34. Dr. Sean Myles, Ph.D. (1/2008—7/2010), Associate Professor, Dalhousie University
35. Dr. Badri Padhukasahasram, Ph. D. (9/2006—1/2008), Bioinformatics Scientist, Illumina, Inc.
36. Dr. Alice Popejoy, Ph.D. (11/2017 – 10/2021), Assistant Professor, U. of California Davis
37. Dr. Morten Rasmussen, Ph.D., (8/2013—7/2016), Director of Technology, ArcBio, LLC
38. Dr. Arindam RoyChoudhury, Ph. D. (7/2008—8/2009), Assistant Professor, Columbia University
39. Dr. Anna Rychkova, Ph.D. (9/2013—04/2016), Scientist, Amgen Comparative Biology and Safety Sciences Department
40. Dr. Suyash Shringarpure, Ph.D. (9/2012—06/2016), Statistical Geneticist at 23andme
41. Dr. Karla Sandoval Mendoza, Ph.D. (4/2010—12/2014), Postdoctoral Scholar, Langebio (Mexico)
42. Dr. Martin Sikora, Ph. D. (7/2010—6/2014), Associate Professor and Group Leader, Center for GeoGenetics, Copenhagen, Denmark
43. Dr. Elena Sorokin, Ph. D. (9/2015—2/28), Statistical geneticist at Ancestry DNA
44. Dr. Scott Williamson, Ph. D. (6/2003—6/2006), deceased (was Assistant Professor at Cornell University)
45. Dr. Genevieve Wojcik, Ph. D. (1/2014—9/2019), Assistant Professor, John Hopkins School of Public Health
46. Dr. Keyan (Kevin) Zhao, Ph.D. (9/2007—12/2010), Staff Scientist, IonTorrent
47. Dr. Ashley Zehnder, DVM, Ph.D. (10/2016—8/2018), Co-founder and CEO at Fauna Bio

Graduate Students Trained

1. Abra Brisbin (Ph. D. Applied Math) (2006 – 2010, joint with J. Mezey), Assistant Professor of Mathematics, University of Wisconsin-Eau Claire
2. Kasia Bryc (Ph. D. Biometry) (2006 — 2010), Population Geneticist, 23andMe
3. Thomas Cooke (Ph. D. Genetics) (2011—2017), Postdoc at Whitehead Institute
4. Helio Costa (Ph. D. Genetics) (2011—2015), Fellow in Molecular Clinical Genetics at Stanford
5. Roxana Daneshjou (M.D./Ph.D. Genetics) (2012—2014, joint w/ Russ Altman), MSTP student at Stanford, completing medical training
6. Joe Davis (Ph. D. Genetics) (2012—2017, joint with S. Montgomery), Data Scientist at Cardinal Analytix Solutions

7. Jeremiah Degenhardt (Ph. D. Biometry) (2005 — 2010), Senior Scientist in Bioinformatics, Onyx Pharmaceuticals
8. Kirsten Eilertson (Ph. D. Statistics) (2007—2011), Assistant Professor, Penn State
9. Hong Gao (Ph. D. Genetics) (2004—2008), Senior Bioinformatics Scientist, Illumina, Inc.
10. Alex Han (Ph.D. Genetics) (2014—2020)
11. Ryan Hernandez (Ph. D. Biometry) (2003—2008), Associate Professor, UCSF
12. Julian Homburger (Ph. D. Genetics) (2014—2017, joint with E. A. Ashley), Data Scientist at Color Genomics
13. Alex Ioannidis (Ph.D. Computational and Mathematical Engineering) (2013—2018), Postdoctoral scholar, Stanford
14. Sandra Indarraga (M.S./Ph. D. Genetics) (2009—2011), self-employed
15. Emilia Huerta-Sanchez (Ph. D. Applied Math, joint with R. Durrett) (2005—2009), Assistant Professor, UC Merced
16. Jeffrey Jensen (Ph. D. Genetics & Development, joint with C.F. Aquadro) (2002—2006), Head of Laboratory and Professor, Ecole Polytechnique Federale de Laussane
17. Lin Lin (Ph. D. Computational Biology) (2006—2010), Postdoctoral Scholar, Harvard University
18. Kirk Lohmueller (Ph. D. Genetics & Development) (2005 – 2009, joint with A. G. Clark), Assistant Professor, UCLA
19. Xin (Maria) Ma (Ph. D. Statistics) (2008—2011), Postdoctoral Scholar, Stanford University
20. Brian Maples (Ph.D. Biomedical Informatics) (2011–2014), Computational Data Scientist, Nuna, Inc.
21. Alicia Martin (Ph.D. Genetics) (2011—2015), Assistant Professor, Mass General Hospital, Harvard University
22. Kimberly McManus (M.S. Genetics) (2012—2015), Data Scientist, LinkedIn
23. Shaila Musharoff (Ph.D. Genetics) (2009—2015), Assistant Professor, Cornell University
24. Patricia Ortiz-Tello (M.D./Ph.D. Genetics) (2010—2016), independent
25. Armin Pourshafeie (Ph.D. Physics) (2015—2019), Citadel
26. David Poznik (Ph.D. Biomedical Informatics) (2011—2015), Computational Biologist and Population Geneticist, 23andme
27. Jessica Torres (Ph.D. Biomedical Informatics) (2014—2021)
28. Mark Wright (Ph.D. Genetics & Development) (2007—2010), 23andMe
29. Fouad Zakharia (Ph.D. Genetics) (2010—2012), Vice President, Morgan Stanley
30. Lan Zhu (M.S. /Ph.D. Biometry) (2002—2006), Associate Professor (tenured), Oklahoma State University

Courses Taught

BIODS210: Configuration of the Health Care System and the Application of Big Data/Analytics (F2016)

BTRY 407: Principles of Probability and Theoretical Statistics (F2007)

BTRY 694: Statistical inference in Population Genomics (S2004,S2006, F2006, F2007, S2008, F2010)

BTRY 484/684: Statistical Genomics (F2005, F2006)

BTRY 694: Statistical Inference in Evolutionary Genomics (S2005)

BTRY 384: Intro. Bioinformatics and Statistical Genomics (F2004, F2005)

BTRY 652: Computationally Intensive Statistical Inference (S2003)

CSRE 110: Comparative Studies of Race and Ethnicity (W2011, guest lecture)

GENE 210: Personalized Genomics (S2010, S2011, guest lecture)

GENE 203: Graduate Genetics (S2013, F2010, lectures on human genetics and genomics)

GENE 203: Graduate Genetics (F2011, primary lecturer)

GENE 202: Genetics for Medical Students (F2010, lectures on human population genetics)
GENE 209: Population Genetics (S2011, S2012, S2013, S2014)

Invited Conferences, Keynote Lectures, and Departmental Seminars (>175 total)

2024

1. CHDI's 19th Annual Huntington's Disease Therapeutic Conference, Feb. 29, 2024; Palm Spring, CA)
2. JP Morgan Biotechnology Conference, Feb. 22, 2024; Taipei, Taiwan

2023

1. ABRCMS (Annual Biomedical Research Conference for Minoritized Scientists) 2023 Keynote Speaker, Nov. 11, 2023; Phoenix, AZ
2. F/Prime, Foresite Capital, Pivotal Biosciences AI in Healthcare Conference 2023, Nov. 6, 2023; San Francisco, CA
3. AI and Quantum Computing, Nov. 21, 2023; Miami, FL
4. WCPG (World Congress in Psychiatric Genetics), Oct. 11, 2023; 2023 Keynote Speaker, Montreal, CA
5. Regeneron Research Center Keynote Speaker, Sept. 18, 2023; Terrytown, NY
6. Fidelity Biosciences Research Institute Post-Acute Infection Syndrome (PAIS) 2023 Conference, Aug. 4, 2023; Newport, RI
7. Colombia Congress on Human Genetics Keynote Speaker, June 15, 2023; Medellin, Colombia
8. Gruter Institute Conference 2023, April 18, 2023; Lake Tahoe, NV
9. Stanford University Biostatistics Workshop, Sept. 28, 2023; Stanford, CA

2022

1. F/Prime Conference on AI and Healthcare, Nov. 14, 2022; London, UK
2. Cornell University, Nov. 4, 2022; Ithaca, NY
3. Harvard Medical School, Sept. 20, 2022; Boston, MA
4. Genetic Health Information Network Summit, Sept 14, 2022; Nashville, TN
5. Yale University, New Haven, CT
6. Fidelity Biosciences Research Institute Post-Acute Infection Syndrome (PAIS) 2022 Conference, Aug. 11, 2022; Newport, RI
7. Pharmacogenetics Research Network (PGRN 2022) University of Colorado, Aurora, CO
8. University of California, San Francisco, San Francisco, CA
9. Gruter Foundation, Mar 22, 2022; London, UK
10. Executive War College, April 29, 2022; New Orleans, LA
11. Post-Covid Summit 2022, April 19, 2022; Brussels, Belgium
12. Tec de Monterrey, Feb 2-4, 2022; Monterrey, Mexico

2021

1. University of Miami, Board of Trustees
2. Harvard University, Cambridge, MA
3. Stanford University Medical Grand Rounds, Stanford, CA
4. Johnson and Johnson, Inc.
5. Cambridge University, Cambridge, UK

2020

1. E.E. Just Lecture University of Chicago, Chicago, IL

2. Stanford University School of Law, Stanford, CA

2019

1. Etalon Diagnostic Leadership Retreat, Santa Rosa, CA
2. Helix Leadership Retreat, Menlo Park, CA
3. DRIFT Symposium, American Society of Human Genetics 2019, Houston, TX
4. NHGRI Workshop: Perspectives in Comparative Genomics & Evolution, Washington, D.C.
5. Human Genetics and Genomics Gordon Research Conference 2019, Closing Keynote, Waterville, NH
6. Stanford Executive Summer Program, Stanford University Graduate School of Business
7. STAT+ Inaugural Event on Venture Capital in Pharma and Biotechnology, San Francisco, CA
8. RECOMB 2019 Keynote Speaker, Washington, D.C.
9. California Institute of Regenerative Medicine (CIRM) Conference on Brainstorming Neurodegeneration” Keynote, San Francisco, CA
10. Nation Health Care Council (NHCC) Fellow’s Program 2019, Nashville, TN
11. Personalized Medicine World Conference 2019, Santa Clara, CA

2018

1. Advances in Genome Biology and Technology Personalized Medicine Keynote, San Diego, CA
2. Genomic Medicine XI, San Diego, CA
3. Genetic Health Information Network Summit, Nashville, TN
4. Siriraj Hospital Genomics and Medicine Keynote, Bangkok, Thailand
5. NHGRI Town Hall, Stanford, CA
6. 2018 Marshall Scholars’ Forum, Stanford, CA
7. Distributed18 (Blockchain in Healthcare) Conference, San Francisco, CA
8. Big Data in Biomedicine 18 (Blockchain organizer), Stanford, CA
9. Stanford Executive Summer Program, Stanford University Graduate School of Business
10. Broad Institute of Harvard/MIT Medical and Population Genomics, Cambridge, MA
11. Frontiers in Genomic Medicine Keynote, Sioux Falls, SD
12. American College of Medical Genetics Annual Meeting Keynote, Charlotte, NC
13. Vanderbilt University Medical Center Personalized Medicine Keynote, Nashville, TN
14. Colorado State University on Veterinary Electronic Health Records, Ft. Collins, CO
15. Pacific Symposium on Biocomputing (PSB 2018) Keynote, Hilo, HI

2017

1. Health 2.0 Keynote Speaker, San Francisco, CA
2. 2017 Distinguished Presidential Scholar, University of Miami, FL
3. CSHL’s Second Annual Double Helix Day, NY, NY
4. Stanford Symposium for Genomics & Personalized Medicine, Menlo Park, CA
5. Stanford Executive Summer Program, Stanford University Graduate School of Business
6. Big Data in Biomedicine Conference 2017, Stanford, CA
7. Stanford-China Cardiovascular Research and Medicine Symposium, Stanford, CA
8. FH 2017 Global Summit, Miami, FL
9. Genetic Health Information Network Summit, Nashville, TN
10. IMSS XXVI National Health Research Forum, San Luis Potosí, Mexico

2016

1. 2016 Darwin Day Keynote Speaker, University of Iowa, Iowa City, IA
2. University of Iowa Medical School Distinguished Speaker, Iowa City, IA

3. University of Iowa Department of Biology Seminar, Iowa City, IA
4. 2nd Annual NIH-wide Pi Day Speaker, National Institutes of Health, Bethesda, MD
5. Personalized Medicine World Conference (PMWC 2016), Silicon Valley, CA
6. Rockefeller University Friday Seminar Speaker, New York, NY
7. UCSF Parnassus Institute for Human Genetics Seminar, San Francisco, CA
8. UCB Supernetwork Conference, London, UK
9. New York City Medical Examiner's Office, NY, NY
10. SAP Sapphire 2016, Orlando, FL
11. Stanford Institute for Computational and Mathematical Engineering (ICME) Expo, Stanford, CA
12. Miami Children's Hospital Speaker Series, Miami, FL
13. Stanford Pediatric Grand Rounds and Retreat Keynote Speaker, Stanford, CA
14. Big Data in Biomedicine Conference 2016, Stanford, CA
15. 2016 CARDIA Conference Keynote Speaker, Napa, CA
16. Perspectives in Genomics Speaker, Cancun, Mexico
17. UCLA Institute for Pure and Applied Mathematics Seminar, Los Angeles, CA
18. Immunogenomics 2016, HudsonAlpha Institute, Huntsville, AL
19. Baylor College of Medicine Molecular & Human Genetics Seminar, Houston, TX

2015

1. Cornell University Department of Biological Statistics and Computational Biology Seminar, Ithaca, NY
2. Festival of Genomics, San Francisco, CA
3. Human Genome Variation and Complex Genome Analysis (HGV 2015) Conference, San Francisco, CA
4. Member, Iconiq Capital Patients Like You Panel, San Francisco, CA
5. Invited lecture at San Francisco Mexican Museum, San Francisco, CA
6. Gordon Conference on Human Genetics and Genomics Speaker and Vice Chair, Newport, RI
7. Inaugural lecture on "Genomics and Health Disparities" at the National Institutes of Health, Bethesda, MD
8. Graduate Student invited Speaker, Department of Human Genetics, University of Chicago, Chicago, IL
9. Keynote speaker, UCLA/IPAM "Latinos in Mathematics" conference, Los Angeles, CA
10. Panelist at "Big Data, Genomics, and Precision Medicine," Tech Museum, San Jose, CA
11. University of Chicago Department of Genetics Seminar, Chicago, IL
12. Keynote speaker, "Human Genome Meeting 2015", Kuala Lumpur, Malaysia
13. Advances in Genome Biology and Technology (AGBT), Marco Island, FL
14. "Genomics and Sequencing Data Integration, Analysis and Visualization" Symposium, San Francisco, CA
15. CIENCIAS—a forum for presenting science in Spanish, Stanford University, Stanford, CA
16. Plant and Animal Genome Conference, San Diego, CA

2014

1. TEDMED 2014 inaugural cohost and invited speaker, San Francisco, CA
2. New York Genome Center Seminar, New York, NY
3. MacArthur Foundation Fellows Forum, Chicago, IL
4. Why We Can't Wait: Conference to Eliminate Health Disparities, Chair and Speaker, Washington D.C.
5. Human Genome Meeting on "Genome Variation and Human Health" (HUGO 2014), Geneva, Switzerland

6. Universidad Nacional del Altiplano, Puno, Peru
7. Illumina Users Meeting, South San Francisco, CA
8. "1000 Genomes and Beyond" Community Meeting, Cambridge, UK
9. Thermo Fisher Next Generation Sequencing seminar, South San Francisco, CA
10. University of Minnesota / Illumina "Understanding Your Genome", Minneapolis, MN
11. SAP Sapphire, Orlando, CA
12. Harvard Medical School Department of Genetics, Student-Invited Speaker, Boston, MA
13. Broad Institute of Harvard and MIT Medical and Population Genomics Seminar, Cambridge, MA
14. Advances in Genome Biology and Technology (AGBT) 2014, Marco Island, FL
15. University of California Riverside Department of Biology, Riverside, CA

2013

1. University of Miami Department of Human Genetics, Miami, FL
2. Willamette University Department of Biology, Portland, CA
3. Gordon Research Conference on Human Genetics and Genomics, RI
4. Arizona State University, MTBI program, Phoenix, AZ
5. Society for Mathematical Biology (SMB) Keynote Speaker 2013, Phoenix, AZ
6. 9th Molecular Biology of Hearing and Deafness Conference, Stanford, CA
7. 2013 Genome 10K Meeting Invited Speaker, Ft. Lauderdale, FL
8. Stanford/Oxford Big Data Conference, Stanford, CA

2012

1. Memorial Sloan Kettering Presidential Invited lecture
2. Oxford-Stanford Conference on Big Data, Oxford, UK
3. Illumina Users Meeting, American Society of Human Genetics 2012
4. Royal Society 2012 Galton and Fisher Lecture invited speaker, London, UK
5. 50th Anniversary of NIGMS DeWitt Stetten Jr. Lecture
6. UCSF Parnassus Seminar, San Francisco, CA
7. University of Helsinki Conference on Next-Generation Sequence Analysis
8. 2012 Green Mountain Conference on Forensic DNA Keynote Speaker
9. 2012 US/Indonesia Kavli Symposium (Solo, Indonesia)
10. University of Oxford Wellcome Trust Center for Human Genomics, Oxford, UK
11. Why We Can't Wait: 2012 Conference on Genomics and Health Disparities, Miami, FL
12. Boston College Landmarks in Biology Seminar, Boston, MA
13. North Carolina State Department of Genetics, Raleigh, NC
14. University of North Carolina Conference on Genetics of the Peoples of Africa and the Transatlantic African Diaspora, Chapel Hill, NC
15. University of California Davis, Davis, CA
16. University of California Berkeley, Berkeley, CA
17. Stanford University School of Law, Stanford, CA

2011

1. Institute for Pure and Applied Math Keynote Speaker, Los Angeles, CA
2. Society for Advancement of Chicanos and Native Americans in Science (SACNAS) 2011 conference, San Jose, CA
3. International Congress of Human Genetics 2011 Satellite Conference on Populations of the New World
4. University of California, Santa Cruz
5. University of Washington, Genome Sciences

6. 3rd EMBO Conference on Host Genetics and Infectious Disease, Pasteur Institute
7. Human Genome Variation (HGV) Meeting 2011
8. American Genetic Association Meeting 2011
9. European Society of Human Genetics 2011
10. University of Oregon (student invited speaker)
11. MacArthur Fellows Meeting
12. Stanford University Medical Grand Rounds
13. Slim Initiative in Genomic Medicine in the Americas
14. Commonwealth Club of California
15. Broad Institute of Harvard/MIT (Population and Medical Genetics)
16. Yale University School of Medicine (Genetics)
17. Harvard University (Organismic and Evolutionary Biology)
18. Florida International University (Biology and Human Genetics)
19. 2011 Genome 10K Conference
20. 2011 Gordon Conference on Quantitative Genomics
21. U. California Berkeley (Computational Biology)
22. Advances in Genome Biology and Technology (AGBT)

2010

1. U. California San Francisco (Human Genetics)
2. National Science Foundation Invited Speaker
3. LifeTech Users Meeting 2011
4. Cold Spring Harbor Banbury Conference on "DNA and Human History"
5. CARTA Symposium on Evolution of Human Biodiversity (UCSD)
6. Princeton University Lewis Sigler Institute
7. UCLA Human Genetics Retreat Invited Speaker
8. Johns Hopkins University School of Medicine (Genetics)
9. Genetic Society of America (GSA) 2010 Model Organism Conference invited speaker
10. European Molecular Biology Lab (EMBL) Workshop
11. Institute Pasteur, Paris, France (Immunology)
12. Duke University Computational Biology Graduate Student Invited Speaker
13. Advances in Genome Biology and Technology (AGBT) 2010)
14. University of Alabama ALLELE Invited Speaker
15. Graduate University for Advanced Studies (Hayama, Japan)

2009

1. Plant and Animal Genome XVI Meeting Aquaculture Keynote speaker
2. Office of AIDS Research Workshop on "Genomic Analyses of HIV Infection"
3. Conservation Genetics (ConGen) Workshop Smithsonian Tropical Research Institute
4. University of Pennsylvania (Genetics)
5. University of Pennsylvania (Veterinary School)
6. University of Miami Miller School of Medicine (Human Genetics)
7. Mexican National Institute of Genomic Medicine (INMEGEN)
8. Inter American Development Bank (IADB) 50th Annual Meeting (Medellin, Colombia)
9. Broad Institute of Harvard and MIT Medical and Population Genetics Seminar Series
10. University of Massachusetts at Amherst (Biology)
11. University of Miami Miller School of Medicine, HIV Grand Rounds
12. University of Washington St. Louis School of Medicine and UW Genome Center
13. RECOMB 2009 Symposium Speaker

14. Stanford University School of Medicine (Genetics)
15. 74th Cold Spring Harbor Symposia on Quantitative Biology
16. Pritzker School of Medicine University of Chicago (Human Genetics)
17. Society for Molecular Biology and Evolution 2009 Meeting (Iowa State University)
18. Gordon Research Conference on Human Genomics invited speaker
19. NGS2009 Conference on Next Generation Sequencing: Challenges and Opportunities (Barcelona, Spain)
20. American Society of Human Genetics (ASHG) 2009 Special Symposium Speaker
21. Columbia University Computational Biology Seminar Series
22. University of Indiana and University of Oregon NSF IGERT Invited Speaker
23. PABEW: Second Biological Evolution Workshop in (Porto Alegre, Brazil)
24. National Academy of Sciences Sackler Symposium "In the light of Evolution IV"

2008

1. Plant and Animal Genome XVI Meeting
2. University of California Los Angeles (History and Language Seminar Series)
3. University of Washington (Genome Sciences)
4. Gordon Conference on Molecular Evolution
5. Cornell University (Baker Institute for Animal Health)
6. University of Michigan (Biostatistics)
7. Harvard University (Anthropology)
8. Cornell University Plant Breeding and Genetics)
9. Stanford University (Biology)
10. Third International Conference on Primate Genomics
11. Cold Spring Harbor Biology of Genomes Meeting
12. 70th Anniversary Celebration for Centro Nacional de Investigacion en Cafe (CENICAFE) Bogota (Colombia)
13. Miami Institute of Human Genomics
14. HGM 2008: HUGO's 13th Human Genome Meeting in Hyderabad, India
15. Banbury Conference on Kinship, Race, and Ancestry
16. Cold Spring Harbor Symposium on Personalized Genomics
17. NHGRI Workshop on "Ethics and the Study of Natural Selection"
18. American Society of Human Genetics (ASHG) 2008 Invited Speaker
19. Stanford University (Biology)
20. 23andMe
21. Cornell-Weill School of Medicine (Pediatric Grand Rounds)

2007

1. Recomb Satellite Workshop
2. University of Arizona Ecology and Evolutionary Biology
3. American Institute of Biological Sciences Annual Meeting
4. NIH Workshop on Macaque Genetics
5. Society for Molecular Biology and Evolution 2007 Annual Conference
6. Banbury Conference on Domestication Genomics
7. National Association of Biology Teachers (NABT)
8. Konrad Lorenz Institute Workshop on Selective Sweeps

2006

1. Institute for Pure and Applied Math Short Program
2. University of Chicago
3. Universitat Pompeu Fabra
4. Arizona State University
5. Society for Molecular Biology and Evolution 2006 Annual Conference)
6. SNP Meetings 2006
7. American Society of Human Genetics ASHG 2006
8. Cornell Center for Vertebrate Genomics

2005

1. Eastern Great Lakes Molecular Evolution Meeting invited speaker
2. Mathematical Biosciences Institute Conference on Evolutionary Genomics
3. Hunter College, invited speaker

2004

1. Max Planck Institute for Evolutionary Anthropology
2. Cornell University (Field of Genetics)
3. North Carolina State (Bioinformatics)
4. Cornell University (Plant Biology)
5. University of Southern California (Computational Biology)
6. University of California Los Angeles (Human Genetics)
7. NSF Institute for Pure and Applied Mathematics
8. ETH Zurich (Student Invited Speaker)
9. University of Bergen
10. Blackwell-Tapia Conference
11. University of Maryland (Ecology & Evolutionary Biology)

2001-2003

1. University California Los Angeles (Biology)
2. University California Los Angeles (Human Genetics)
3. University of Arizona (Biology)
4. California Institute of Technology (Biology)
5. University North Carolina Chapel Hill (Genetics)
6. Brown University (Biology)
7. Cornell University (Biometry)
8. University of California Berkeley