# Application of High-Throughput Technologies to Genomic Analysis of Polygenic Traits

by

Ernest Tsz-Tsun Lam

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

# Application of High-Throughput Technologies to Genomic Analysis of Polygenic Traits

by

Ernest Tsz-Tsun Lam

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Pui-Yan Kwok Chair

STEVEN HAMILTON

John S. Witte John S. Witte

Committee in Charge

# Acknowledgements

This dissertation represents a body of work that many have contributed to, and there are countless people that I must thank. I offer my most sincere gratitude to those that have helped in one way or another and those that have lent their support in making this possible.

First and foremost, I would like to thank *Prof. Pui-Yan Kwok* for being my mentor and graduate advisor. He took me in when I knew little about genetics (let alone genomics) and research in general. He has entrusted me with the optical mapping project that he has been so passionate about, and for that, I am extremely grateful. He is always supportive. He has given me the freedom and time I needed to grow and develop as an independent scientist. Most importantly, he has created an environment that fosters collaboration instead of competition. He has brought in so many genuinely nice people to this lab, and this has had an unquantifiable positive impact on my graduate work.

I am grateful to past and present members of the Kwok Lab; it is an honor to have worked with all of you. In this age of Big Science, one could not conceal the fact that most of our work requires extensive collaboration. Our lab members are our most immediate collaborators. *Angel Mak* – thank you for your honesty and thoughtfulness. You have helped me see and appreciate things with your childlike sense of wonder. You should believe in yourself, that you *truly* are special in your own ways. *Annie Poon* – I could not have gotten here without your help with the CREATE project and the optical mapping project. I have enjoyed your company. Thank you for your generosity that I wish I could repay. I am grateful for the breakfast and the soup that you brought me. I am grateful for the countless rides I have gotten from you, many of which not always the most convenient for you. *Bani Tamraz* – I thank you for always going the extra mile (quite literally). He drove us to visit Illumina in Hayward; we went to Gladstone together. You are always generous and welcoming. *Blanca Herrera* – I enjoyed working briefly with you on the kidney exome project. I wish you, your soon-to-come baby, and your family all the best. *Catherine Chu* – thank you for keeping me healthy by buying me chicken wings, driving us to badminton, and bringing me delicious soup. I am really glad I got to know Alex as well. *Jennifer Pons* – thank you for letting me play with your gene expression data and for welcoming my participation. *Ludmila Pawlikowska* – you were almost my second mentor. I remember going to you sometimes late in the afternoon when I wanted to talk, and you never turned me away. I thank you for listening to me and for giving constructive suggestions. I would not have gotten past my early years without you. *Nina Gonzaludo* – I have always admired your entrepreneurial spirit. Thanks for sharing your industry experience with us. *Paul Ling-Fung Tang* – I am now an Apple believer because of you. I really appreciate your openness and willingness to explore. I have enjoyed our conversations. I sincerely hope that you will be able to find something that you are passionate about. *Stacy Musone* – thank you for taking on the role of the senior student/mentor. You were always so nice and kind. Thanks for listening to me when I needed someone that would understand. *Stephanie Hesselson* – thank you for organizing our lab events and for all the treats. I had a lot of fun playing with Etta. I will always remember the hamburger that you made when I visited. *Ting-Fung Chan* – thank you for your generosity and hospitality when I visited your lab in Hong Kong. I appreciate your mentorship, and I hope that we will continue to collaborate. *Wilson Liao* – thank you for being my mentor when I first rotated in the lab. Thank you for brining me to the psoriasis clinic. The experience was life-changing for me. I really enjoyed working with you on the ZNF750 project.

Over the past few years, I have had the opportunity to work with many talented people: *Yang Cao*, *Emma Dowd*, *Aparna Chhibber*, *Jesus Lopez*, *Dean Ehrlich*, *Dillon Dong*, and *Jasmin Eshragh*. Every summer, the lab frantically decides what projects would be appropriate for our summer students. Mentoring can be draining at times, but there is only joy in being able to share my passion for science with them. I will readily admit that I am not the best mentor despite my genuine intentions, but they have tolerated my shortcomings and appreciated my efforts. Most importantly, all of them have gone on to bigger things. As a mentor, I cannot be happier for them. They have inspired me, pushed me, and brought me joy.

I thank my thesis committee members – *Prof. Pui-Yan Kwok*, *Prof. Steve Hamilton*, and *Prof. John Witte* – for their guidance and support. In particular, I have had extended interactions with Steve because of our collaboration on the Border collie deafness project. I am eternally grateful that *Jennifer Yokoyama* brought me onto the project. We met at Starbucks, laid out a plan, got all the reagents, prepared all the samples… I could not have asked for a better collaborator. And of course, I thank her for her friendship. Her infectious laughs. Her honesty. Her selflessness. I am very excited about our project, which has been fruitful, and working with Jennifer and Steve could not have been more wonderful.

I would like to thank my qualifying exam committee members – *Prof. Nadav Ahituv*, *Prof. Xin Chen*, *Prof. Kathy Giacomini*, and *Prof. Hao Li* – for their valuable input. They asked difficult questions not to make me stumble but to force me to think about broader impacts of my research. *Nadav* and *Sook Wah* Yee graciously invited me to participate in the review paper in Genome Medicine, and I am honored to have worked with them.

I would like to thank *Prof. Xin Chen*, *Prof. Neil Shah*, and *Prof. Charles Craik* for taking me as a rotation student. Among other things, *Cindy Kosinski* helped with my experiments and my NSF application, and I really appreciate her friendship. *Corynn Kasap* was particularly patient with me when teaching me cell culture techniques. *Christopher Weier* taught me how to run Western blots and stayed late with me when I needed to test antibody combinations. He was fun to be around with and always so kind.

I have had the fortune of participating in multiple collaborative projects, and several of those have been fruitful. I have the most sincere respect for *Prof. James Cleaver*, who introduced me to the XP proteins. I felt intimidated when I was told you are a member of the Academy of Sciences. You took the time to talk to me about more than just science. Thank you for the career advice you have given me. I can’t thank you enough for inviting me to participate in the Nature Reviews Genetics paper.

*Ming Xiao* played a huge role in the optical mapping project. Thank you for doing almost daily calls with us in order to help set up TIRF when you didn’t need to. I have enjoyed our dinner conversations, and I have learned so much from you. Our optical mapping work is a work of true collaboration. I am proud of our product. I hope you are too.

*Greg Tranah* was another one of my main collaborators. You welcomed my participation when I expressed interest. You introduced me to the wonderful world of mitochondrial genetics. You were always positive, engaging, and welcoming.

*Prof. Michael Zwick* welcomed me to spend a week with his lab, and *David Okou* shared his protocol for solid-phase target capture with me. David taught me how to QC, what to look for, and everything else about the protocol. Everyone at the Zwick Lab was very kind to me.

*Niranjan Nagarajan* has provided much needed input on alignment of optical maps. His experience in this area of research has helped the project tremendously. I have enjoyed our weekly Skype sessions.

I also want to thank my early mentors. *David Lauter* (George Washington High School) was the one that got me first started. I wouldn't have gotten to where I am if I didn't take his Biology and AP Biology classes. He was a very special teacher, and I hope that others are inspired by him in the same way he inspired me to pursue science.

I started doing research in college because of the Chem 91 Undergraduate Honors Seminar at UC San Diego. *Prof. Jack Dixon* was the speaker for one of the evenings. I vaguely remember he talked about kinases and phosphatases and cancer. He encouraged us to try research, and when I got back to my dorm room, I immediately emailed him. We talked on the phone on the next day, and he offered me a volunteer position in his lab. I thank *Shaun Lee* for being patient with me when I did not know *anything* (literally) about research. It wasn't until much later that I understood what NRPS really meant. I then worked with *Prof. Michael VanNieuwenhze* in the Chemistry and Biochemistry Department. *Aaron Wohlrab* taught me the basics of synthetic chemistry. I saw first-hand how incredibly hardworking organic chemists were.

I have enjoyed interacting with my classmates from the PSPG graduate program. We are a group with diverse interests, but everyone is kind and supportive. They even bought me my first beer at PSPG225 Research Opportunities/Pizza Talk when I turned 21. It is because of the little Perl script *Joseph Lee* wrote for me that I started learning Perl. We have lived together since the beginning of graduate school. I got a glimpse of some of his cooking (molecular gastronomy) adventures, his bitcoin mining operation (modern alchemy), and his trading strategies. He always has cool ideas that I wish I had thought of.

My friends outside of my academic bubble have been an indispensable source of support. They have kept me sane and grounded. I want to thank *Hannah Yeung* and *Heidi Yuen* for cheering me up with their occasional whatsapp messages and for spending time with me when I visited Hong Kong. Hannah, thanks for sending me all the pictures of the amazing foods that you made. I have always appreciated your sense of optimism. Heidi, I treasure our friendship because of the ups and downs that we have been through, despite our differences. C*han(dler) Ho*, my roommate in college, for his friendship. I am glad we have kept in touch all these years. He has been a good friend and practically a member of the family.

***Lastly and most importantly***, I would like to thank my family for being there for me. I am blessed with their unwavering support and unconditional love. They gave me my Chinese name, Tsz-Tsun, which roughly translates to “strive for perfection”. It is perhaps the source of my perfectionistic tendencies. They taught me to always be kind and generous. They instilled in me a sense of wonder, curiosity. I thank you for your trust in me. I thank you for *Everything*.

### *Research funding and support*

E.T.L was supported in part by the NIH Pharmaceutical Sciences and Pharmacogenomics Training Grant (T32 GM007175) and the UCSF Graduate Dean's Health Science Fellowships.

### *Original manuscript citations and co-authorship*

Portions of the text of this dissertation are reprints of material as it appears in the following published or submitted manuscripts:

1. ***Lam ET***, Bracci PM, Holly EA, Chu C, Poon A, Wan E, White K, Kwok P-Y, Pawlikowska L, Tranah GJ. Mitochondrial DNA sequence variation and risk of pancreatic cancer. *Cancer Research* (2012). Feb 1;72(3):686-95.

2. Yokoyama JS*, ***Lam ET****, Ruhe AL, Erdman CA, Robertson KR, Webb A, Williams DC, Chang ML, Lohi H, Hamilton SP, Neff MW. Variation in genes related to cochlear biology is strongly associated with adult-onset deafness in Border collies. *Under revision at PLoS Genetics.* (*Contributed equally)

3. Smith RP*, ***Lam ET****, Markova S, Yee SW, Ahituv N. Pharmacogene regulatory element discovery using next-generation sequencing technologies. *Genome Medicine* (2012). May 25;4(5):45. (*Contributed equally)

4. Tranah GJ, ***Lam ET***, Katzman SM, Nalls MA, Zhao Y, Evans DS, Yokoyama JS, Pawlikowska L, Kwok P-Y, Mooney S, Kritchevsky S, Goodpaster BH, Newman AB, Harris TB, Manini TM, Cummings SR for the Health, Aging and Body Composition Study. Mitochondrial DNA sequence variation is associated with free-living activity energy expenditure in the elderly. *In press at Biochimica et Biophysica Acta – Bioenergetics.*

5. Tranah GJ, Nalls MA, Katzman SM, Yokoyama JS, ***Lam ET***, Zhao Y, Mooney S, Thomas F, Newman AB, Liu Y, Cummings SR, Harris TB, Yaffe K for the Health, Aging and Body Composition Study. Mitochondrial DNA sequence variation associated with dementia and cognitive function in the elderly. *In press at Journal of Alzheimer's Disease.*

6. ***Lam ET***, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok P-Y. Nano-mapping for structural variation analysis and sequence assembly. *In press at Nature Biotechnology.*

# Application of High-Throughput Technologies to Genomic Analysis of Polygenic Traits

Ernest T Lam

## Abstract

Understanding DNA sequence variation is the first step to understanding the underlying genetic architecture of a complex trait, often a manifestation of joint contributions from multiple genes. While instrumental to the successful completion of the Human Genome Project and still considered to be the gold standard, Sanger sequencing is rapidly being phased out for discovery as newer technologies that deliver sequence data with higher throughput and at lower costs become available. With the rise of these newer technologies, generation of sequencing data is no longer the bottleneck. However, one is faced with having to decide between competing technologies, and interpretation of sequencing data remains a significant challenge. In this dissertation, I will present my research on three sequencing projects and one genome mapping project. The first sequencing project involves the use of an array-based resequencing platform, the Affymetrix MitoChip to resequence the mitochondrial DNA for a pancreatic cancer case-control study and for a subset of samples from the Health ABC study. Array-based resequencing represents a high-throughput option for resequencing selected, well-defined regions. We uncovered many novel, rare variants in the mitochondrial sequence, some of which are highly conserved, protein-altering, and predicted to be damaging. While their functional consequences are not clear, we show that more rare variants remain to be found, and the abundance of them suggests that mitochondrial variants could contribute to disease phenotypes. Then, two targeted sequencing studies will be described. The first used an emulsion PCR based approach to study variants in genes involved in drug pathways; the second used a solution-phase target capture approach to fine map causative variants in a region in canine chromosome 6 associated with adult-onset deafness in Border collies. Finally, I will describe our use of nanochannel arrays for genome mapping, an improved version of optical mapping, in structural variation analysis and sequence assembly. Sequence motif maps were constructed 95 BACs previously sequenced for the MHC Sequencing Consortium. The BAC set was sequenced and assembled *de novo* with assistance from the sequence motif scaffold. We show that the sequence motif map could serve as a scaffold, facilitating gap closing and assembly finishing. Together, our work highlights the potential roles these new technologies play in genetic studies and the importance of taking a multigene approach while considering polygenic traits.

**Chapters**

*Chapter 1 – Introduction*

*Chapter 2 – Role of mitochondrial DNA variants in cancer and ageing*

*Chapter 3 – Discovery of variants underlying chemotherapy response by emulsion PCR-based targeted sequencing of chemotherapy pathway genes*

*Chapter 4 – GWAS-guided solution-phase target capture and mapping of adult-onset deafness in Border collies*

*Chapter 5 – Nanomapping of the MHC region: structural variation analysis and de novo sequence assembly*

*Chapter 6 – Final comments*

## Dissertation Organization

This dissertation is organized as follows. ***Chapter 1*** gives a general summary of recent developments in high throughput sequence analysis. It serves to review how technologies have driven new discoveries. I aim to highlight the strengths and weaknesses of various technologies being applied in genetic studies. In ***Chapter 2***, I describe the application of the Affymetrix MitoChip platform for large-scale resequencing of the mitochondrial genome. We explore the potential role of mitochondrial DNA variants in the context of cancer and ageing, as motivated by the hypothesis that mitochondrial dysfunction is associated with these phenotypes. In ***Chapter 3***, I describe the use of multiplex emulsion PCR and next generation sequencing to identify variants associated with extreme chemotherapy response phenotypes. In ***Chapter 4***, I describe fine-mapping of a locus associated with adult-onset deafness in Border collies in a genome-wide association study using solution-phase target capture and next generation sequencing. After extensive filtering based on conservation, segregation patterns, and properties of the variants, putative variants of interest are identified and validated in additional samples via Sanger-based genotyping. ***Chapter 5*** explores sequence motif mapping as a potential tool for structural variation analysis and *de novo* sequence assembly. As proof of concept, we used 95 BACs that tile the MHC region from two BAC libraries and showed that a high quality sequence motif map scaffold could provide long-range information useful for *de novo* assembly and assembly finishing. In ***Chapter 6***, I conclude with some final comments about the future of high-throughput genomics and some of the outstanding challenges of the field.

## Table of Contents

**Chapter 1: Introduction**

**Chapter 2: Role of mitochondrial DNA variants in cancer and ageing**

**Chapter 5 Nanomapping of the MHC region: structural variation analysis and *de novo* sequence assembly**

**Chapter 6 Concluding remarks and future directions**

## List of Tables

## List of Figures

# Chapter 1

## Introduction

Development of genomic tools is essential as they benefit a wide range of disciplines in genetics such as disease mapping, comparative and evolutionary genomics and population genetic. These tools have enabled genome-scale studies of gross chromosomal organization and abnormalities at one extreme and studies of single-base sequence variation at the other. Benefiting from these new tools' unprecedented throughput, we hope to gain a more comprehensive understanding of sequence variation in healthy and diseased individuals.

### *1.1 The beginning*

Chain-terminator sequencing, or Sanger sequencing as pioneered by Frederick Sanger and colleagues [1] in the 1970s is one of the earliest methods enabling efficient DNA sequencing. Central to the method is the use of dideoxynucleotides that get incorporated in the DNA but prevent further synthesis because of a modified functional group. In the classical method, there are four separate sequencing reactions, each containing one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, and ddTTP) and unmodified nucleotides. These dideoxynucleotides lack a 3'-OH group, inhibiting proper formation of the phosphodiester bond. As the dideoxynucleotides get incorporate in a random fashion, fragments of varied lengths are generated. The DNA products are separated via gel electrophoresis for each reaction. The DNA sequence can be then "read" based on the relative positions of the bands in the four lanes. Later developments include incorporation of an automated optical system, capillary electrophoresis, and fluorescently labeled ddNTPs. These have enable automated DNA sequencing. Sequencing reads of up to around 1000 bases could be easily obtained using Sanger sequencing; however, the first and last bases are often of lower quality.

There are several factors limiting the maximum read length one can achieve with Sanger sequencing. First, separating large DNA fragments in a capillary with one-base differences is difficult. While there is a 10% difference in length between a 10 bp fragment and a 11 bp fragment; the difference between a 1000 bp fragment and a 1001 bp fragment is much more subtle. Second, the chain-terminating dideoxynucleotides are incorporated based on a binomial probability. It is more likely that the chain is terminated at some earlier point than for example, after 1000 bases. Therefore, the fluorescent intensities tend to degrade towards the end of a read, leading to high base calling error rates. Third, longer fragments spend more time traversing the gel matrix than short fragments. The base-calling accuracy decreases as the corresponding peaks tend to be wider and their widths more variable.

Considerations such as peak shape and resolution are included in the calculation of the Phred quality score, which is assigned to each base call to denote the probability of a given being called incorrectly [2, 3]. Phred scores have been useful for deriving consensus sequences, guiding removal of low-quality segments of a sequence read, and providing a general assessment of the sequence quality. Formally, a Phred quality score $Q$ is defined as being logarithmically related to the base-calling error probability $P$ in the following manner: $Q = -10 \log_{10}P$.

### *1.2 The Human Genome Project and the "$1000 Genome" era*

An international collaborative effort, the Human Genome Project (HGP) [4] was one of the most important landmarks in human genetics, and it demonstrated the capability of Sanger sequencing. While deemed highly successful in terms of laying the groundwork for future studies, it was a costly undertaking, and parts of the human genome were poorly covered or assembled. Genome sequencing using the Sanger method requires cloning of DNA fragments into plasmids of different sizes such as yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), and cosmids. The fragments have to be compatible with the vector carrier. As such, there are regions that could not be cloned. Even though Sanger sequencing currently provides the longest reads, complete *de novo* assembly is challenging without the help of physical maps. Repetitive regions are particularly difficult to assemble because the reads are not sufficiently long enough to span the repeats. Repeat-filled centromeres and heterochromatic regions important for proper chromosome segregation and gene expression regulation are relatively sparsely covered [5]. Resolving these regions requires extensive manual assembly finishing using carefully selected tiling BACs and other physical and cytogenetic map information [6], but computational approaches are showing promise as well [7].

Since the completion of the HGP, it is well-recognized that routine genome sequencing using the Sanger method is not economically possible for most labs. If the sequencing cost is not significant lowered, genome sequencing would only be feasible for large genome centers. In recent years, there has been a push to lower the cost of genome sequencing of an individual to under $1,000. George Church, a vocal proponent for personal genome sequencing, posits that as sequencing becomes more affordable, individuals will be more open to the once-in-a-lifetime expenditure of having one's own genome sequenced, eventually realizing the promises of personalized medicine [8]. To promote advancement, the National Institutes of Health (NIH) awarded more than $38 millions in grants to development of new sequencing technologies (http://www.genome.gov/12513210) in 2004. These include "Near-Term Development for Genome Sequencing" projects that would enable sequencing of the human genome for less than $100,000 within five years and "Revolutionary Genome Sequencing Technologies" projects that take on the longer-term challenge of sequencing human genome for under $1,000. On the private sector side, the J. Craig Venter Science Foundation and the X Prize Foundation have offered the $10 million Archon X Prize, for team(s) "able to sequence 100 human genomes within 30 days to an accuracy of one error per 1,000,000 bases, with 98% completeness, identification of insertions, deletions and rearrangements, and a complete haplotype, at an audited total cost of $1,000 per genome [9-11]." This "$1000 Genome" movement have prompted multiple sequencing approaches to be explored as discussed in the following sections.

### *1.3 Sequencing-by-hybridization*

The concept and theory of sequencing-by-hybridization were first proposed in the late 1980s [12, 13]. One attractive feature is its departure from the reliance of polymerases and ligases. Sequencing-by-hybridization can be summarized in two steps. First, oligonucleotide probes of known sequences are hybridized to the DNA target one intends to sequence. Then, one can identify all the N-mers that are present in the target and reconstruct the complete sequence. In the original manuscript, Drmanac *et al.*

estimated that 95,000 11-nucleotide probes would be needed to sequence one million bases [12]. It was not until the advent of solid-phase microarrays that the vision was realized [14, 15]. Since then, array-based platforms have been developed and commercialized by companies such as Affymetrix and Illumina for genotyping and (re)sequencing applications. Because millions of oligonucleotide probes can now be synthesized and immobilized onto a solid-support array, highly parallel interrogation of DNA sequences became possible. The input DNA can either be PCR products or fragmented whole-genome amplified DNA. Typically, the DNA template of interest is incubated at an elevated temperature with fluorescently labeled oligonucleotide probes tiled on a microarray. Hybridization is detected via automated imaging. The probe containing the appropriate complementary sequence to the target DNA would give the brightest signal. The original sequence can then be reconstructed computationally. For resequencing of a particular region, probes are tiled with 1 bp offsets such that each base can be interrogated.

Microarray-based sequencing is highly flexible and well-suited for resequencing of genomic regions of interest that can be targeted by probes. Kothiyal *et al*. described an array design that included 13 genes implicated in sensorineural hearing loss (SNHL) [16] and explored the potential clinical utility of resequencing arrays. The authors found that with an average call rate of 99.6% and accuracy of >99.8%, the arrays could facilitate quicker and cheaper diagnosis over serial molecular techniques, and they complement traditional non-genetic laboratory and radiographic tests.

Array resequencing has found various applications in pathogen genomics as reviewed in [17]. It is useful for typing, comparing, and epidemiological tracking strains of the same pathogen or related species. An Affymetrix resequencing array was custom-designed to sequence 29,212 base-pairs (which corresponds to 0.5% of the *B. anthracis* genome) in *B. anthracis* [18]. Based on resequencing of 56 *B. anthracis* strains, it was concluded that the average level of sequence variation was low and variable sites were less likely to be protein-altering. In a more recent report, 303 kb were sequenced in each of 39 *B. anthracis* strains from the Biological Defense Research Directorate's collection [19]. The denser resequencing allowed phylogenetic analyses of the strains and confirmed the low level of sequence variation previously found. Broad-spectrum detection of a variety of bacteria and DNA and RNA viruses was made possible by tiling conserved regions of various species on a single resequencing array [20]. This allows for simultaneous unbiased testing for the presence of multiple pathogens. There was good concordance between conventional analyses (culture and/or PCR/RT-PCR) and the array approach. However, it was noted that this approach may not work as well for species that are mutating quickly. An array was designed against the 29.7-kb SARS coronavirus genome using a maskless array synthesis process such that new sequence variation can be quickly added as needed [21]. Should there be another outbreak, the array could be used for high-throughput and cost-effective investigation of any new strains.

A resequencing array was designed to target the entire mitochondrial genome and commercialized by Affymetrix (under the name "Affymetrix GeneChip Mitochondrial Resequencing Array", or "MitoChip" in short) in a single assay. It provides higher throughput than methods relying on amplification of the mitochondrial DNA with overlapping PCR amplicons followed by Sanger sequencing [22]. In its first design, the MitoChip (v1.0) contained probes that target both strands of the entire human mitochondrial coding sequence, with the majority of the probe arrayed in duplicates [23]. In the second iteration (v2.0; also the current version), additional probes were designed specifically to target the noncoding regions and common single-nucleotide variants, insertions and deletions [24]. The MitoChip has been independently validated using 93 worldwide samples [25]. A comparison between the calls made by the MitoChip and by traditional Sanger sequencing revealed that the MitoChip achieved an average call rate of 99.48% and an accuracy of >99.98% using default base calling settings. The base call rate and sequence accuracy can be further improved using custom software [26-28].

Various studies have used the MitoChip for studying disease phenotypes, most notably cancer. Dasgupta *et al*. found the same clonal mtDNA mutations in both histologically normal airway mucosal biopsies and the paired lung tumor samples [29]. This study and others [30, 31] suggest that comprehensive mtDNA resequencing could be a diagnostic tool for detecting mutations with potential association with cancer.

The MitoChip could also be useful in the context of forensic testing, when there is insufficient genome DNA for autosomal typing because of the high abundance of mitochondrial DNA relative to nuclear DNA [32]. Complete mitochondrial sequencing could provide more information than traditional forensic typing of just the hypervariable regions.

Sequencing-by-hybridization is not without limitations. Because the target DNA is compared to complementary probes of the array, one is restricted to studying the contents included in the array design. Also, *de novo* assembly and sequencing of large regions or genomes would be difficult as hybridization is error-prone (because of cross-hybridization and in the presence of sequence variation) and repeat structures are hard to de-convolute.

### *1.4 Next generation sequencing*

In the International HapMap project, millions of single nucleotide polymorphisms (SNPs) were identified and their allele frequencies characterized in worldwide populations [33]. Since the inception of the project, SNPs have become the preferred type of genetic marker. They are broadly distributed across the genome with different allele frequencies and provide a denser genetic map than for example satellite markers. The use of microarrays has further enabled high throughput simultaneous genotyping of millions of SNPs genome-wide. Many common SNPs (typically referring to those with minor allele frequencies greater than 5%) and regions in linkage disequilibrium with them have been shown statistically associated with common disease phenotypes and complex traits through genome-wide association studies (GWAS). The NHGRI maintains a catalog of published GWAS (http://www.genome.gov/gwastudies/), and as of May 25, 2012, the database includes 1,269 publications, implicating 6,439 strongly supported SNPs in a wide range of phenotypes. However, most associated SNPs explain only a small portion of the heritability of a given phenotype [34]. Although the heritability estimates have been disputed and argued to be over-estimated [35], there has been a general shift away from focusing on common variants to rare variants and structural variants in hopes of identifying additional contributing factors. The growing interest in sequencing is precisely fueled by this realization that all types of variants need to be considered in understanding complex traits, and that rare variants can only be comprehensively surveyed by sequencing.

Because of the aforementioned limitations, sequencing-by-hybridization has largely been replaced by a group of methods collectively termed next-generation sequencing (NGS). NGS is usually associated with sequencing technologies commercialized by companies such as Illumina (Genome Analyzer, and later, HiSeq) and Life Technologies (SOLiD and 454). Sample preparation often starts with fragmentation of the DNA to around 300-500 base pairs, followed by end-repair. Platform-specific universal adapters are added to the ends of the fragments to facilitate ligation-mediated amplification. With differences in the enzymes used and the exact sequencing chemistry, they generally rely on sequential incorporation and fluorescent imaging of labeled nucleotides in single molecule-derived clusters of template DNA either attached to beads or a solid support (discussed in more detail in [36]). Several human genome sequences have been made available because of the dramatic decrease in sequencing cost [37-39]. The publically funded 1000 Genomes Project whose goal is to sequence more than 1,000 complete human genomes using NGS technologies is expected undercover the majority of the remaining common variants and a large fraction of rare variants [40].

In the beginning, the Illumina and SOLiD platforms could sequence one end of the fragments for only around 30 bases. It is now possible to sequence both ends of the fragments for around 100 bases. Different NGS platforms provide comparable sensitivity in variant discovery and genotyping accuracy given high sequencing coverage. Each technology has its own error characteristics. 454 pyrosequencing is known to have issues dealing with homopolymers, but empirical distributions and 454-specific error characteristics have been derived based on the raw flow value data [41]. For Illumina sequencing, at high coverage, errors tend to be systematic and are dependent on local sequence contexts [42]. Sanger sequencing remains the gold standard for validation of variants found using NGS technologies because of its longer reads and higher accuracy.

Despite its higher per-base cost compared to the Illumina platform, the 454 platform boasts read lengths of more than 200 bases and has been used in metagenomic applications [43]. Simulated 450 bp pyrosequencing reads with a 0.49% error rate gave rise to higher metagenomic gene prediction error rates while Sanger reads, albeit not perfect, were shown better suited for gene prediction [44]. Nonetheless, 454 sequencing complements Sanger sequencing by filling gaps from hard stops [45], where the trace signal drops dramatically in regions of the traces that should be resolvable. These hard stops are often encountered when there is potential secondary structure (for example, hairpins and stem-loops) in the DNA template or when the region being sequenced has a high %GC content. Because the DNA is randomly fragmented (often via sonication) during the sample preparation process for most NGS protocols, formation of secondary structure is reduced. However, coverage might still be uneven across regions with extreme %GC contents.

### *1.5 Targeted sequencing*

Even though the throughput has increased and the costs reduced dramatically since the shift from Sanger sequencing to NGS technologies, whole-genome sequencing of a large number of individuals is still relatively costly. Therefore, candidate gene or targeted sequencing approaches become an attractive alternative. Campbell *et al.*

developed a PCR-based approach to amplify the clone-specific VDJ rearrangement at the Ig heavy chain locus (IGH) found to be hypermutated in B-cell chronic lymphocytic leukemia cells (CLL) [46]. The PCR products were sequenced using the Roche 454 platform, and the high-coverage sequencing revealed subclonal structures of cancer cells and rare clones, enabling reconstruction of the clonal evolution of the cancer cell populations. PCR amplification remains a popular choice for enriching for specific regions of the genome. It was used to amplify all protein-coding exons in various cancers [47-49], and the PCR products were subjected to Sanger sequencing. The authors were able to characterize variants in the coding sequences and identify driver mutations. PCR has been proven effective for amplifying single regions as in the example of the CLL study, but large scale PCR amplification of a large number of regions is time-consuming and highly laborious. For example, the above-mentioned cancer studies required hundreds of thousands of PCR reactions.

RainDance Technologies commercializes an emulsion PCR approach that is able to amplify hundreds and thousands of regions in parallel. In this method, a microfluidic chip is used to merge droplets containing fragmented template DNA with droplets containing single pairs of PCR primers. Millions of droplets containing single PCR reactions are merged and collected. After PCR, the emulsion is broken and the PCR products are combined for sequencing. In the proof-of-concept study, Tewhey *et al.* demonstrated successful coupling of microdroplet-based emulsion PCR of 3,976 amplicons and NGS to sequence 435 exons of 47 genes encompassing 1.49 Mb of sequence [50]. In terms of coverage and genotyping accuracy, this multiplex PCR approach was shown to be as effective as the conventional PCR approach.

Hybridization-based target capture has also been explored. Okou *et al*. [51] and Hodges *et al*. [52] first proposed hybridizing the template DNA to probes targeting sequences of interest tiled on a microarray. After hybridization, the unwanted and unbound DNA can be washed away. Gnirke *et al*. [53] later showed that solution-phase capture, which is similar in concept to array-based solid-phase capture but involves hybridization of the template DNA to probes in solution, provides better capture performance because of more favorable hybridization kinetics and is more amenable to automation because most of the steps can be performed in plates. Currently, a single solution-phase capture design can target more than 50 Mb of sequence. Coupled with NGS, "exome" designs provide a focused view of high-value regions of the genome believed to have the highest *a priori* probabilities of harboring functional variants. The first demonstration described in Ng *et al*. [54] and as shown in numerous subsequent studies [55-59], exome sequencing has been instrumental in identifying the causative variants in disease phenotypes.

There have been studies comparing the performance of recently developed target capture methods [60, 61]. It is generally concluded and reflected in its popularity that solution-phase capture is currently the most effective method for targeting a large set of sequences. Targeted sequencing will remain important for hypothesis-driven studies, and for the time being, exome sequencing represents a cost-effective and informative alternative to whole-genome sequencing.

### *1.6 Optical mapping*

Besides single-nucleotide changes, structural variation is also a class of genetic variation of interest. The term structural variation encompasses copy-number variation, insertions, deletions, duplications, inversions, and other rearrangement events. The consideration of structural variants for the missing heritability is natural – these variants could affect large stretches of DNA (as compared to single-base changes) and are found to be relatively common in modern populations. One early example was found in Charcot-Marie-Tooth disease, which was associated with a duplication in chromosome 17p [62]. More recent studies have highlighted the potential of structural variants to disrupt neurological functions, as demonstrated in [63, 64].

Traditional cytogenetic studies have shown that there are large structural variants; however, the resolution is limited because only very large variants can be visualized using microscope-based methods. Microarrays initially used for genotyping have been adopted for structural variation typing and discovery. For example, one can identify a deletion by noting a drop in the fluorescent signal across a region relative to the neighboring regions. Studies can be performed in a high throughput fashion with low costs; however, microarrays have their limitations. Microarrays are most effective for discovery of large deletions. Copy number neutral and balanced structural variation cannot be comprehensively surveyed using existing methods. While copy-number variants have been deposited in the Database of Genomic Variants and Database of Genomic Structural Variants at an exponential rate, there has been a minimal increase in the number of entries for inversions [65].

Next generation sequencing data have been used to structural variation discovery at a single-base resolution. There are several ways to look for structural variation using

NGS data. By taking advantage of the paired-read information, one can identify read pairs that map anomaly. One can analyze the distribution of read depth to identify regions of excess and depleted coverage. One can also look at instances where a single read can be split and mapped two distinct regions in the genome in order to identify for example the breakpoint of a deletion [66, 67]. However, Onishi-Seebacher and Korbel [67] suggested that there is ascertainment bias in using short reads for analyzing structural variants. For example, complex structural variants are difficult to decipher and are obscured in short reads. Therefore, there is a need to overcome these challenges in order to comprehensively catalogue structural variants.

First developed by David Schwartz and colleagues in the 1990s [68], optical mapping provides a physical map that has proven to be useful for structural variation analysis [69] and assembly of microbial [70, 71] and eukaryotic genomes [72-74]. In this method, large single DNA molecules are cleaved at sequence-specific sites by a restriction enzyme. The DNA backbone is stained and the fragments visualized via microscopy, resulting in an ordered, genome-wide physical map. Because large DNA molecules of hundreds of thousands of base pairs are used as template, long-range relationships between loci are preserved in contrast to short-read NGS technologies. A deletion can be identified by simply noting where there are missing fragments as compared to an *in-silico* map derived from the reference sequence. Optical mapping has been fully automated and commercialized by OpGen, and further developments include the use of microfluidic devices that provide higher throughput [75]. Nicking enzymes that introduce single-stranded breaks are used so that the DNA can be nicked, labeled, and imaged inside micro- or nano-fluidic devices [76].

Optical mapping could also have potential in assisting sequence and genome assembly. It can validate and order sequencing contigs and help with genome finishing [77]. With adequate coverage, *de novo* assembly of the sequence data is possible, and it could eliminate the need for comparison with the reference sequence. Much research has been devoted to exploring the possibility of *de novo* assembly of short reads. One major appeal of *de novo* assembly is the independent from an often-incomplete reference sequence. Algorithmic improvements have led to improved assemblies [78, 79]; however, current approaches are not sufficient. It is expected that the incorporation of high-resolution physical map data (such as those from optical mapping) would provide a useful scaffold for *de novo* assembly of short reads. Because of the long-range information imbedded in the optical maps, the scaffold could potentially reduce the complexity of *de novo* assembly.

### *1.7 Future developments*

Third-generation sequencing platforms are soon coming (reviewed in detail in [80]). Besides promising throughput and longer read lengths, they have the advantage that amplification of the DNA template is no longer needed because of the single-molecule nature of these newer approaches. It is expected that some of the non-linear biases will be minimized, simplifying downstream analyses. Several studies have explored the possibility of using nanopores for sequencing [81]. Most approaches involve driving single-stranded DNA or RNA molecules through nanopores and monitoring resulting changes in the current as each nucleotide passes through. This type of single-molecule nanopore technologies offers several advantages over current methods – no amplification is needed, and the cost is further cut down by the elimination of polymerases, ligases or nucleotides during sequencing. However, to achieve single-base resolution for practical use, further research is needed to optimize the transverse speed and the channel length.

In terms of structural variation and haplotype analysis, it is likely that with longer read lengths and further computational advancement, sequencing will become the *de facto* technology of choice. The new Illumina MiSeq platform will likely enable paired-end sequence of 200 bases and more; the Pacific Biosciences platform could generate reads of thousands of bases [82]. The longer read lengths will extend our capability to survey structural variation of different sizes. On the computational side, software programs like BEAGLE [83, 84] for performing phasing and imputation for large datasets continue to be in active development. As mentioned in the above section, optical mapping and other high-resolution physical mapping technique could also provide important long-range information. By combining sequencing and high-resolution physical mapping, a complete picture of the genome could emerge with all single-nucleotide variants and structural variation phased and annotated.

### *1.8 References*

1. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
2. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.
3. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
4. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

5. Eichler, E.E., R.A. Clark, and X. She, *An assessment of the sequence gaps: unfinished business in a finished human genome.* Nat Rev Genet, 2004. **5**(5): p. 345-54.

6. Hoskins, R.A., et al., *Sequence finishing and mapping of Drosophila melanogaster heterochromatin.* Science, 2007. **316**(5831): p. 1625-8.

7. Lee, H.R., K.E. Hayden, and H.F. Willard, *Organization and molecular evolution of CENP-A--associated satellite DNA families in a basal primate genome.* Genome Biol Evol, 2011. **3**: p. 1136-49.

8. Church, G.M., *Genomes for all.* Scientific American, 2006. **294**(1): p. 46-54.

9. Kedes, L. and E.T. Liu, *The Archon Genomics X PRIZE for whole human genome sequencing.* Nat Genet, 2010. **42**(11): p. 917-8.

10. Kedes, L. and G. Campany, *The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition.* Nat Genet, 2011. **43**(11): p. 1055-8.

11. Kedes, L., et al., *Judging the Archon Genomics X PRIZE for whole human genome sequencing.* Nat Genet, 2011. **43**(3): p. 175.

12. Drmanac, R., et al., *Sequencing of megabase plus DNA by hybridization: theory of the method.* Genomics, 1989. **4**(2): p. 114-28.

13. Drmanac, R., et al., *Sequencing by hybridization (SBH): advantages, achievements, and opportunities.* Adv Biochem Eng Biotechnol, 2002. **77**: p. 75-101.

14. Ramsay, G., *DNA chips: state-of-the art.* Nat Biotechnol, 1998. **16**(1): p. 40-4.

15. Warrington, J.A., et al., *New developments in high-throughput resequencing and variation detection using high density microarrays.* Hum Mutat, 2002. **19**(4): p. 402-409.

16. Kothiyal, P., et al., *High-throughput detection of mutations responsible for childhood hearing loss using resequencing microarrays.* BMC Biotechnol, 2010. **10**: p. 10.

17. Steinberg, K.M., D.T. Okou, and M.E. Zwick, *Applying rapid genome sequencing technologies to characterize pathogen genomes.* Anal Chem, 2008. **80**(3): p. 520-8.

18. Zwick, M.E., et al., *Microarray-based resequencing of multiple Bacillus anthracis isolates.* Genome Biol, 2005. **6**(1): p. R10.

19. Zwick, M.E., et al., *Genetic variation and linkage disequilibrium in Bacillus anthracis.* Sci Rep, 2011. **1**: p. 169.

20. Lin, B., et al., *Application of broad-spectrum, sequence-based pathogen identification in an urban population.* PLoS One, 2007. **2**(5): p. e419.

21. Wong, C.W., et al., *Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays.* Genome Res, 2004. **14**(3): p. 398-405.

22. Jakupciak, J.P., et al., *Mitochondrial DNA as a cancer biomarker.* Journal of Molecular Diagnostics, 2005. **7**(2): p. 258-267.

23. Maitra, A., et al., *The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection.* Genome Res, 2004. **14**(5): p. 812-9.

24. Zhou, S., et al., *An oligonucleotide microarray for high-throughput sequencing of the mitochondrial genome.* Journal of Molecular Diagnostics, 2006. **8**(4): p. 476-82.
25. Hartmann, A., et al., *Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes.* Hum Mutat, 2009. **30**(1): p. 115-22.
26. Thieme, M., et al., *ReseqChip: Automated integration of multiple local context probe data from the MitoChip array in mitochondrial DNA sequence assembly.* Bmc Bioinformatics, 2009. **10**.
27. Zamzami, M.A., et al., *Insights into N-calls of mitochondrial DNA sequencing using MitoChip v2.0.* BMC Res Notes, 2011. **4**: p. 426.
28. Xie, H.M., et al., *Mitochondrial genome sequence analysis: a custom bioinformatics pipeline substantially improves Affymetrix MitoChip v2.0 call rate and accuracy.* Bmc Bioinformatics, 2011. **12**: p. 402.
29. Dasgupta, S., et al., *Following mitochondrial footprints through a long mucosal path to lung cancer.* PLoS One, 2009. **4**(8): p. e6533.
30. Jakupciak, J.P., et al., *Performance of mitochondrial DNA mutations detecting early stage cancer.* BMC Cancer, 2008. **8**: p. 285.
31. Mithani, S.K., et al., *Mitochondrial resequencing arrays detect tumor-specific mutations in salivary rinses of patients with head and neck cancer.* Clin Cancer Res, 2007. **13**(24): p. 7335-40.
32. Vallone, P.M., J.P. Jakupciak, and M.D. Coble, *Forensic application of the Affymetrix human mitochondrial resequencing array.* Forensic Sci Int Genet, 2007. **1**(2): p. 196-8.
33. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
34. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
35. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.
36. Metzker, M.L., *Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation.* Nature Reviews Genetics, 2010. **11**(1): p. 31-46.
37. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.
38. Wang, J., et al., *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-5.
39. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
40. *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
41. Balzer, S., et al., *Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim.* Bioinformatics, 2010. **26**(18): p. i420-5.
42. Harismendy, O., et al., *Evaluation of next generation sequencing platforms for population targeted sequencing studies.* Genome Biol, 2009. **10**(3): p. R32.
43. MacLean, D., J.D. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nat Rev Microbiol, 2009. **7**(4): p. 287-96.

44. Hoff, K.J., *The effect of sequencing errors on metagenomic gene prediction.* BMC Genomics, 2009. **10**: p. 520.

45. Goldberg, S.M., et al., *A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes.* Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11240-5.

46. Campbell, P.J., et al., *Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing.* Proc Natl Acad Sci U S A, 2008. **105**(35): p. 13081-6.

47. Parsons, D.W., et al., *An integrated genomic analysis of human glioblastoma multiforme.* Science, 2008. **321**(5897): p. 1807-12.

48. Jones, S., et al., *Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.* Science, 2008. **321**(5897): p. 1801-6.

49. Parsons, D.W., et al., *The genetic landscape of the childhood cancer medulloblastoma.* Science, 2011. **331**(6016): p. 435-9.

50. Tewhey, R., et al., *Microdroplet-based PCR enrichment for large-scale targeted sequencing.* Nat Biotechnol, 2009. **27**(11): p. 1025-31.

51. Okou, D.T., et al., *Microarray-based genomic selection for high-throughput resequencing.* Nat Methods, 2007. **4**(11): p. 907-9.

52. Hodges, E., et al., *Genome-wide in situ exon capture for selective resequencing.* Nat Genet, 2007. **39**(12): p. 1522-7.

53. Gnirke, A., et al., *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.* Nat Biotechnol, 2009. **27**(2): p. 182-9.

54. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.

55. Bilguvar, K., et al., *Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations.* Nature, 2010. **467**(7312): p. 207-10.

56. Puente, X.S., et al., *Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome.* Am J Hum Genet, 2011. **88**(5): p. 650-6.

57. Shi, Y., et al., *Exome sequencing identifies ZNF644 mutations in high myopia.* PLoS Genet, 2011. **7**(6): p. e1002084.

58. Caputo, V., et al., *A restricted spectrum of mutations in the SMAD4 tumor-suppressor gene underlies Myhre syndrome.* Am J Hum Genet, 2012. **90**(1): p. 161-9.

59. Li, M., et al., *Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar ataxia in a Chinese family.* Clin Genet, 2012.

60. Teer, J.K., et al., *Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing.* Genome Res, 2010. **20**(10): p. 1420-31.

61. Kiialainen, A., et al., *Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery.* PLoS One, 2011. **6**(2): p. e16486.

62. Lupski, J.R., et al., *DNA duplication associated with Charcot-Marie-Tooth disease type 1A.* Cell, 1991. **66**(2): p. 219-32.

63. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-43.

64. Diskin, S.J., et al., *Copy number variation at 1q21.1 associated with neuroblastoma.* Nature, 2009. **459**(7249): p. 987-91.
65. Baker, M., *Structural variation: the genome's hidden architecture.* Nat Methods, 2012. **9**(2): p. 133-7.
66. Koboldt, D.C., et al., *Massively parallel sequencing approaches for characterization of structural variation.* Methods Mol Biol, 2012. **838**: p. 369-84.
67. Onishi-Seebacher, M. and J.O. Korbel, *Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond.* Bioessays, 2011. **33**(11): p. 840-50.
68. Schwartz, D.C., et al., *Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping.* Science, 1993. **262**(5130): p. 110-4.
69. Teague, B., et al., *High-resolution human genome structure by single-molecule analysis.* Proc Natl Acad Sci U S A, 2010. **107**(24): p. 10848-53.
70. Lai, Z., et al., *A shotgun optical map of the entire Plasmodium falciparum genome.* Nat Genet, 1999. **23**(3): p. 309-13.
71. Lin, J., et al., *Whole-genome shotgun optical mapping of Deinococcus radiodurans.* Science, 1999. **285**(5433): p. 1558-62.
72. Church, D.M., et al., *Lineage-specific biology revealed by a finished genome assembly of the mouse.* PLoS Biol, 2009. **7**(5): p. e1000112.
73. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
74. Zhou, S., et al., *Validation of rice genome sequence by optical mapping.* BMC Genomics, 2007. **8**: p. 278.
75. Jo, K., et al., *A single-molecule barcoding system using nanoslits for DNA analysis.* Proc Natl Acad Sci U S A, 2007. **104**(8): p. 2673-2678.
76. Zhang, P., et al., *Engineering BspQI nicking enzymes and application of N.BspQI in DNA labeling and production of single-strand DNA.* Protein Expr Purif, 2010. **69**(2): p. 226-34.
77. Zhou, S., et al., *A single molecule scaffold for the maize genome.* PLoS Genet, 2009. **5**(11): p. e1000711.
78. Li, H., *Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly.* Bioinformatics, 2012.
79. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.
80. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
81. Branton, D., et al., *The potential and challenges of nanopore sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1146-53.
82. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
83. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* Am J Hum Genet, 2007. **81**(5): p. 1084-97.

84. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* Am J Hum Genet, 2009. **84**(2): p. 210-23.

# Chapter 2

## Role of mitochondrial DNA variants in cancer and ageing

This chapter contains content from the following three manuscripts:

*(Published at Cancer Research)* **Mitochondrial DNA sequence variation and risk of pancreatic cancer**

Ernest T. Lam, Paige M. Bracci, Elizabeth A. Holly, Catherine Chu, Annie Poon, Eunice Wan, Krystal White, Pui-Yan Kwok, Ludmila Pawlikowska, Gregory J. Tranah

*(In press at Biochim et Biophys Acta - Bioenergetics)* **Mitochondrial DNA sequence variation is associated with free-living activity energy expenditure in the elderly**

Gregory J. Tranah, Ernest T. Lam, Shana M. Katzman, Michael A. Nalls, Yiqiang Zhao, Daniel S. Evans, Jennifer S. Yokoyama, Ludmila Pawlikowska, Pui-Yan Kwok, Sean Mooney, Stephen Kritchevsky, Bret H. Goodpaster, Anne B. Newman, Tamara B. Harris, Todd M. Manini, and Steven R. Cummings, for the Health, Aging and Body Composition Study.

*(In press at Journal of Alzheimer's Disease)* **Mitochondrial DNA sequence variation associated with dementia and cognitive function in the elderly**

Gregory J. Tranah, Michael A. Nalls, Shana M. Katzman, Jennifer S. Yokoyama, Ernest T. Lam, Yiqiang Zhao, Sean Mooney, Fridtjof Thomas, Anne B. Newman, Yongmei Liu, Steven R. Cummings, Tamara B. Harris, and Kristine Yaffe, for the Health, Aging and Body Composition Study.

### *2.1 Introduction*

The mitochondrion is an essential organelle that provides much of the cellular energy in the form of adenosine triphosphate (ATP). Genes involved in mitochondrial assembly, metabolism, growth and reproduction are distributed throughout the nuclear and mitochondrial genomes [1, 2]. This large set of genes includes ~100 nuclear- and mitochondria-encoded polypeptide genes for five oxidative phosphorylation (OXPHOS) complexes [1, 2].

The human mitochondrial genome is a circular double-stranded molecule of 16,569 bp in human. It encodes 13 highly conserved polypeptides for the OXPHOS system and the necessary RNA machinery for their translation within the mitochondria. The mitochondrial genome does not recombine, is maternally inherited [3], and has a unique organization in that its genes lack introns, intergenic sequences, and 5' and 3' noncoding sequences.  Each human cell contains hundreds of mitochondria and thousands of copies of mtDNA with the number of copies being dependent on the cell type and its energy requirements.

The 13 mitochondrial proteins are components of complexes I, III, IV, and V. They are essential to mitochondrial energy production and considered the most functionally important [4]. Therefore, sequence variation within the 13 mtDNA-encoded OXPHOS genes may impact superoxide production and ATP generation efficiency through respiratory chain impairment [5, 6], apoptosis [7], and ATP supply [8].

Because human mtDNA has a mutation rate 10-20 times higher than that of nuclear DNA [9-11] and approximately one-third of sequence variants found in the general population may be functionally important [2], it is likely that most of the mtDNA variation that impacts function is rare in frequency and only accessible by direct sequencing. Also, different loci may exhibit different relationships between allele frequency and functional effect. Some may harbor functional alleles at high frequencies, whereas other may have only rare or private functional variants. Thus, it is necessary to consider the combined effect of all mtDNA mutations on physiological and pathological changes.

The entire mitochondrial genome can be sequenced in a single reaction using the Affymetrix Human Mitochondrial Resequencing Array 2.0 (MitoChip). Tiling 25-mer oligonucleotide probes were designed against the Revised Cambridge Reference Sequence (RCRS) as published by MITOMAP (www.mitomap.org). The MitoChip interrogates both the forward and reverse strands of the mitochondrial genome for a total of ~30 kb sequence. It enables the detection of known and novel mutations. For example, the design includes probes for common variants in the HV1 and HV2 regions chosen from the FBI database. The MitoChip has redundant probes for detecting the major human mitochondrial haplotypes and known disease-related mutations. Built-in redundancy via independent probe sets also allows a test of within-chip reproducibility. Two overlapping long-range PCR reactions can be used to amplify the mitochondrial genome from total DNA samples. For more degraded DNA samples, traditional short amplicon PCR may be used. In the following sections, I will describe our mitochondrial sequencing work using the MitoChip in a pancreatic cancer case-control study and a study looking at dementia and energy expenditure in samples from the Health ABC cohort.

***2.2 Methods – Mitochondrial sequencing***

Since the studies in this chapter all relied on the MitoChip platform, the protocol is described here. Mitochondrial DNA was amplified from total genomic DNA extracted from platelets and sequenced with the MitoChip using the Affmetrix protocol for GeneChip CustomSeq Resequencing Array as previously described [12].

*Mitochondrial DNA amplification*. The entire mitochondrial genome was first amplified in two long-range PCR reactions using LA PCR Kit (Takara Bio U.S.A.). For the Mito1-2 amplicon (9,307 bp), the forward primer was ACATAGCACATTACAGTCAAATCCCTTCTCGTCCC, and the reverse primer was ATTGCTAGGGTGGCGCTTCCAATTAGGTGC. For the Mito3 amplicon (7,814 bp), the forward primer was TCATTTTTATTGCCACAACTAACCTCCTCGGACTC, and the reverse primer was CGTGATGTCTTATTTAAGGGGAACGTGTGGGCTAT.

*Array resequencing*. After amplification, the resulting PCR products were assessed qualitatively by 1% agarose gel electrophoresis and purified using a Clonetech Clean-Up plate (Clonetech). The purified DNA was quantified by PicoGreen and for a subset of the samples, confirmed by NanoDrop measurements. The amplicons were pooled at equi-molar concentrations. Chemical fragmentation was performed and products were confirmed to be in the size range of 20-200 bp by 20% polyacrylamide gel electrophoresis with SYBR Gold staining. The IQ-EX control template, a 7.5 kb plasmid DNA, was used as a positive control. The samples were labeled using TdT and hybridized to the array in a 49$^{o}$C rotating hybridization oven for 16 hours. Finally, streptavidin phycoerythrin (SAPE), and then antibody staining was performed. The microarrays were processed in the GeneChip Fluidic Station and the GeneChip Scanner.

*Sequence data analysis*. Signal intensity data was output for all four nucleotides, permitting quantitative estimates of allelic contribution. The allelic contribution was assessed using the raw data from the individual signal intensities by deriving the ratio of expected allele (REA), which is the log ratio of the raw signal intensity of the a given allele at any site to the average raw signal intensity of the other three alleles. DAT files with raw pixel data were generated and used as input for grid alignment. CEL files generated from DAT files were analyzed in batches using GSEQ. Samples with call rates of less than 95% were discarded. For samples passing initial filtering, ResqMi 1.2 [13] was used for re-analysis of bases originally called as "N" by GSEQ. Analysis was performed using custom Perl scripts. Data was extracted from gene regions as defined by NCBI annotations for the revised Cambridge Reference Sequence (rCRS; NC_012920.1).

### *2.3 Mitochondrial DNA sequence variation and risk of pancreatic cancer*

This study provides a comprehensive assessment of mitochondrial sequence variants and suggests that aggregated common and rare variants and the accumulation of singleton variants contribute to pancreatic cancer risk.

#### *2.3.1 Abstract*

Although the mitochondrial genome exhibits high mutation rates, common mitochondrial DNA (mtDNA) variation has not been consistently associated with pancreatic cancer.

Here, we comprehensively examined mitochondrial genomic variation by sequencing the mtDNA of participants (cases=286, controls=283) in a San Francisco Bay Area pancreatic cancer case-control study. Five common variants were associated with pancreatic cancer at nominal statistical significance ($p<0.05$) with the strongest association being mt5460g in the *ND2* gene (odds ratio (OR)=3.9, 95% confidence interval (CI)=1.5-10; p=0.004) which encodes an A331T substitution. Haplogroup K was nominally associated with reduced pancreatic cancer risk (OR = 0.32, CI=0.13-0.76; p=0.01) when compared with the most common haplogroup, H. A total of 19 haplogroup-specific rare variants yielded nominal statistically significant associations ($p<0.05$) with pancreatic cancer risk, with the majority observed in genes involved in oxidative phosphorylation. A weighted-sum approach was used to identify an aggregate effect of variants in the 22 mitochondrial tRNAs on pancreatic cancer risk (p=0.02). While the burden of singleton variants in the HV2 and 12S RNA regions was three times higher among European haplogroup N cases than controls, the prevalence of singleton variants in ND4 and ND5 was two to three times higher among African haplogroup L cases than in controls. Together, the results of this study provide evidence that aggregated common and rare variants and the accumulation of singleton variants are important contributors to pancreatic cancer risk.

### *2.3.2 Introduction*

Pancreatic cancer is the fourth leading cause of death from cancer among men and women in the United States and was expected to result in 43,140 new cases and 36,800 deaths in 2010 [14]. Due to its aggressive nature and a lack of early detection methods, pancreatic cancer is metastatic in more than 50% of patients at the time of diagnosis and has a 5-year relative survival rate for all stages of less than 6%. The incidence of pancreatic cancer is higher in industrialized countries, and it varies by age, sex, and race with more than 70% of pancreatic cancer cases being diagnosed after the age of 60 [15].

Otto Warburg first hypothesized that cancer might be caused by defects in the mitochondrion based on his observation that tumors actively metabolize glucose and produce excessive lactate in the presence of oxygen (aerobic glycolysis) [16]. Mitochondria produce most of the cellular energy, generate reactive oxygen species (ROS), and regulate apoptosis. The primary site of ROS and free radical production during oxidative respiration is the inner mitochondrial membrane [17-19], where the mitochondrial DNA (mtDNA) resides. If increased mitochondrial ROS production increases cancer risk, then mtDNA mutations that partially inhibit electron transport and increase ROS production might also increase cancer risk. Although a complete elucidation of the "Warburg effect" has not been achieved, several mechanisms have been proposed to explain this phenomenon [20], and cancer cells have been shown to exhibit multiple alterations in mitochondrial content, structure, function, and activity [21-23].

Jones *et al.* [24] sequenced the complete mtDNA in 15 pancreatic cancer cell lines and xenografts and identified somatic mtDNA mutations and novel variants in nearly all samples. Kassauei *et al.* [25] sequenced the mtDNA in 15 primary pancreatic cancers and noneoplastic tissue and  all pancreatic cancers demonstrated at least one somatic mutation with the number of mutations per case ranging from one to 14. Somatic mtDNA mutations present in primary tumors were also detected in pancreatic juice from the same patients [26].

Mitochondrial haplogroups, and in some cases specific nonsynonomous (NS) SNPs, have been correlated with cancer development [27-35] suggesting that individuals who inherit certain variants might be more prone to cancer. Results to date do not support a significant involvement of common inherited mtDNA variation as a risk factor for pancreatic cancer [36, 37]. However, these previous studies did not comprehensively examine sequence level mtDNA variation including rare variants and singletons (variants unique to a single participant). In the present study we sequenced the entire mtDNA genome (~16.5 kb) using the Affymetrix Mitochondrial Resequencing Array 2.0 (MitoChip) to assess the role of common and rare mtDNA sequence variation in pancreatic cancer in nearly 600 case and control participants from a large population-based study of pancreatic cancer in the San Francisco Bay Area.

### *2.3.3 Methods*

*Study participants.* Detailed methods have been published for this population-based case-control study of pancreatic cancer in six San Francisco Bay Area counties [38]. Briefly, cases with primary adenocarcinoma of the exocrine pancreas were identified using cancer registry rapid case ascertainment. Eligible cases were San Francisco Bay Area residents newly diagnosed in 1995-1999, ranging from 21-85 years old. Patient diagnoses were confirmed by participants' physicians and by the Surveillance, Epidemiology, and End Results abstracts that included histologic confirmation of disease. A total of 532 eligible cases (67%) completed the study interview [38]. Control participants were identified using random-digit dial (RDD) and were frequency-matched to cases by county, sex and 5-year age group. A total of 1,701 eligible control participants (67%) completed the study interview [38]. Participants who were on blood thinning medications, had a bleeding disorder, had a portacath in place, or had other contraindications to blood draw were not eligible to participate in the optional laboratory portion of the study. A total of 309 cases and 964 controls were eligible for venipuncture and participated in the optional laboratory portion of the study by providing a blood specimen. Demographic characteristics of the 309 study cases who provided blood samples for analyses were similar to those who did not (p-values >=0.30 for age at diagnosis, sex, white race and Hispanic ethnicity) whereas control participants who provided a blood sample were similar by age and Hispanic ethnicity (all p-values >=0.81) but were more likely to be white (p=0.002) and male (p=0.002). A total of 297 cases and 301 controls were selected from these for mtDNA sequencing based on DNA availability and were frequency-matched by age and sex. A total of 286 pancreatic cancer cases and 283 controls yielded sequence data of sufficient quality for analysis (>95% success rate).

*Interviews*. Detailed in-person interviews were conducted in the homes of the participants or at a location of their choice. Race/ethnicity was based on self-report and was broadly categorized into Caucasian, black/African-American, Asian, Hispanic (black or white), or "other race/ethnicity".

*Quality control*. Twenty samples were sequenced twice for concordance assessment. Laboratory personnel were blinded to QC and case–control status and all 20 QC samples had >98% sequence concordance (the majority of discordant calls resulted from positions successfully called in one but called as “N” in another).

*Association testing*. We analyzed variants from the following categories: common haplogroups and individual variants (minor allele frequency [MAF] ≥5%); rare variants (MAF <5%); and singletons (occurring in a single participant – in the case of haplogroup-specific analyses, singletons are variants occurring in a single participant within the haplogroup). Unconditional logistic regression was used to obtain odds ratios (ORs) as estimates of relative risks (hereafter called risk) and 95% confidence intervals (CIs) for analyses involving haplogroups and common variants. Allele frequencies and rare variants (excluding singletons) were compared between cases and controls using chi-square tests. The major European mitochondrial haplogroups were defined using variants identified from PhyloTree [39] and included subgroups H, V, J, T, U, K, (B, F), and (A, I, W, X, Y). To account for confounding by ancestry, we derived eigenvectors using principal components analysis (PCA) with the complete mtDNA genotype data [40]. All models were adjusted for age in five-year groups, sex and the first six eigenvectors of mitochondrial genetic ancestry derived from principal component analysis. Models examining individual common variants were restricted to haplogroup N. We did not examine associations for haplogroups L and M since the sample sizes are small and the numbers of cases and controls reflect the proportions of African- and Asian-American participants in the study.

*Joint variant testing*. Variants also were grouped and tested jointly to assess the contribution of multiple variants to the pancreatic cancer risk using weighted-sum statistics computed as described in Madsen and Browning [41]. 1000 permutations of case-control status were performed to obtain one-sided p-values, testing the hypothesis that most rare mutations are deleterious and associated with disease status. Variants from genes encoding the four mtDNA-encoded OXPHOS complexes, rRNA, tRNA and hypervariable regions were assessed using the weighted sum method [41] among haplogroup N participants only. All weighted-sums were computed using custom Perl scripts.

*Singleton testing*. The total number of gene or region specific variants was compared between pancreatic cancer cases and controls using Fisher's exact test as was the number of individuals harboring singleton variants unique to cases or controls. Due to the potential for confounding by mtDNA ancestry, all singleton analyses were performed for each major haplogroup L, M, and N. Bonferroni correction for multiple testing took into account the number of haplogroups (n=8) and genes/regions examined for singleton analysis (n=17), and the number of common and rare variants discovered and analyzed.

*Function prediction*. In-silico prediction methods were employed to examine mtDNA nucleotide conservation and the impact of NS coding substitutions on amino acid protein sequence. PhastCons [42] is a hidden Markov model-based method that estimates the probability that each nucleotide belongs to an evolutionary conserved element. Based on a multi-species sequence alignment the method considers the conservation of sites flanking the base of interest when producing base-by-base conservation scores. The phastCons scores range from 0 to1 and represent probabilities of negative selection. PhyloP [43] separately measures conservation at individual nucleotides, ignoring the effects of their neighbors. Also based on a multi-species sequence alignment, this method is more appropriate for evaluating signatures of selection at particular nucleotides. PhyloP scores represent -log p-values under a null hypothesis of neutral evolution. Sites predicted to be conserved are assigned positive scores and sites predicted to be fast-evolving are assigned negative scores. For phastCons and phyloP, a higher value indicates a more conserved position. The effects of NS coding substitutions (amino acid changes) on protein function was assessed using PolyPhen2 [44]. PolyPhen2 predicts the possible impact of an amino acid substitution on the structure and function of a human protein. PolyPhen2 outputs a posterior probability that a mutation is damaging and qualitatively reports it as benign, possibly damaging, or probably damaging.

### *2.3.4 Results*

Sequencing of 16,543 mtDNA bases (positions 12-16,555) from 286 pancreatic cancer cases and 283 controls participants yielded a cumulative total of 2,169 variants including: 66 common variants (MAF ≥5%); 251 low frequency variants (MAF 1-5%); and 1,859 rare variants (MAF <1%) including 1,393 haplogroup-specific (L, M or N) singletons unique to either cases or controls.

Distributions of age, education level, race, ethnicity, sex, smoking status, and major haplogroups for cases and controls are presented in Table 2.1. The case and control participants were similar by white/non-white race (chi-square, p=0.74) and Hispanic ethnicity (chi-square, p=0.58). In general, cases were slightly less educated, more likely to be current smokers, and a greater proportion of them African-American.

Haplogroup distributions largely overlapped with self-identified race for African-Americans (97%) and European-Americans (96%). Self-identified Asian-Americans were distributed between haplogroups M (55%) and N (43%). While major haplogroup M is largely unique to Asia, the minor Asian haplogroups are descended from both major haplogroups M (e.g. haplogroups C, D, and G) and N (haplogroups A, B, and F). Participants that self-identified as Hispanic were distributed between haplogroups N (75%) and M (19%). The results for the Asian and Hispanic participants are not unexpected, since the mtDNA traces the maternal lineage exclusively and may not reflect an admixed genetic or self-identified ancestry.

### *Common haplogroups and individual variants*

Risk of pancreatic cancer among eight European sub-haplogroups is reported in Table 2.2. No haplogroup met statistical significance after adjustment for multiple comparisons (eight haplogroups, critical a=0.006). Carriers of haplogroup K had a nominally reduced pancreatic cancer risk compared with the most common European haplogroup H (odds ratio (OR)=0.32, 95% confidence interval (CI)=0.13-0.76, p=0.01). There also were no individual common variants that met statistical significance after multiple comparisons adjustment (66 common variants detected by sequencing [MAF ≥5%], critical a=0.0008). Of the 66 common variants, five reached nominal statistical significance (P<0.05) and two yielded a strong (statistically non-significant) association with pancreatic cancer risk: mt5460g (p=0.004) and mt1811g (p=0.008). The mt5460g>a variant associated with an increased risk of pancreatic cancer (OR=3.9, 95% (CI)=1.5-10; p=0.004) encodes an A331T substitution in the *ND2* gene. All analyses were adjusted for age, sex and six eigenvectors of mitochondrial genetic ancestry derived from principal component analysis. Restricting the haplogroup and common variant analyses to self-identified white non-Hispanic did not alter the results (data not shown).

***Rare variants***

Of 710 low frequency (MAF 1-5%) and rare variants (MAF <1%) detected by sequencing (excluding singletons), none met statistical significance after adjustment for multiple comparisons (critical a=0.0001). A total of 19 haplogroup-specific variants yielded nominal statistically significant associations (a<0.05) with pancreatic cancer risk (Table 2.3). All were from either haplogroup N or L. Of these, 13 were detected from complex I, III, IV and V coding regions, 2 from the 12S RNA and 4 from the hypervariable (HV) or non-coding regions. Two of the coding region variants resulted in NS substitutions: L555Q (mt14000) in *ND5* and K6N (mt14763) in *CytB*. The L555Q substitution in *ND5* was predicted to be probably damaging by PolyPhen2. The mt14763 variant underlying the K6N substitution shows strong evidence of belonging to a conserved sequence element (PhastCons=1.00, PhyloP=4.47).

***Weighted-sum pooled variant testing***

Multiple variants across the combined 22 mtDNA tRNA regions were statistically significantly associated with pancreatic cancer as determined by permutation of case-control status (one-sided p=0.02). Our results also suggest that there was an excess of variants in Complex III (*CytB*) among pancreatic cancer cases although the effect was not statistically significant (one-sided p=0.06).

***Singletons***

A total of 1,393 singleton mtDNA variants unique to cases or controls were identified across the coding, tRNA, rRNA and HV regions for the three major haplogroups L, M, and N. Two of these genes/regions harbored a significantly higher burden of singleton variants in cases compared with controls after adjustment for multiple comparisons (18 mtDNA genes/regions, critical a=0.003). Specifically, the number of singleton variants among haplogroup N cases vs. controls was statistically significantly higher in the HV 2 region (p=0.006) and nominally higher in the 12S RNA (p=0.03) region. The frequency of HV 2 and 12S RNA singletons was three times higher in cases than controls: HV 2 region, 9% cases and 3% controls; 12S RNA 7% cases and 2% controls (Figure 2.1). In haplogroup L, the number of singletons was nominally higher among cases for the entire mtDNA (p=0.004), and for complexes I (p=0.03) and IV (p=0.05). Haplogroup L cases had a statistically significant greater number of singleton variants in the *ND5* gene (p<0.001) and nominally greater numbers in the *ND4* (p=0.02), *COII* (p=0.04), and *COIII* (p=0.007) genes compared with controls. The frequency of *ND4* and *ND5* singletons was two to three times higher in cases than controls: *ND4*, 41% cases versus 14% controls; *ND5*, 74% cases versus 36% controls (Figure 2.2). In addition, singleton variants were observed among haplogroup L cases only for the *COII* and *COIII* genes. Among haplogroup M participants, cases had statistically significant fewer singleton variants in the 12S RNA (p=0.03) region compared with controls (Figure

2.3). Of the 1,393 singleton variants identified in the three major haplogroups, 625 were located within the four mtDNA-encoded OXPHOS complexes with 221 resulting in NS coding substitutions. An excess of *ND2* NS substitutions was observed among cases compared with controls (p=0.02). Among the complex V genes, *ATP6* and *ATP8*, approximately 25% of NS substitutions in controls were predicted to be non-conserved whereas all NS substitutions in cases were conserved (p=0.02).

### *2.3.5 Discussion*

In this study, we sequenced the entire mtDNA in a large population-based case-control study of pancreatic cancer to examine the role of haplogroups and common genetic variants, rare sequence variants, and singletons. We observed inverse associations with pancreatic cancer risk for participants from European haplogroup K when compared with the most common European haplogroup H. The haplogroup K association was not seen in a previous study of pancreatic cancer that included replication [37]. The low frequency of haplogroup K participants in this study and lack of consistency with the earlier study likely mean that our finding is a false positive result. Common haplogroup N variants (MAF ≥5%) in Complex I (*ND2*), Complex IV (*COIII*), 16S, tRNA, and HV2 genes/regions also yielded nominally significant associations with pancreatic cancer risk. This includes an A331T substitution in the *ND2* gene that was present in 20 cases (8%) and 6 controls (2%).

Efforts to identify genetic factors that contribute to complex phenotypes such as cancer must be sensitive to the ways that genes and genetic perturbations operate. For example, it is now widely recognized that common genetic variants play a much smaller role in mediating phenotypic expression and disease risk than initially thought [45-48] and that identification of causative variants requires comprehensive resequencing of genomic loci in multiple subjects [49]. In this study, we identified numerous low frequency (MAF 1-5%) and rare variants (MAF <1%) from haplogroups N and L that were associated with pancreatic cancer risk. Nearly seventy percent of these variants were observed in the OXPHOS complexes including two NS variants from the *ND5* and *CytB* genes. The *ND5* substitution was predicted to have a damaging effect on the resulting protein whereas the *CytB* substitution showed evidence of belonging to a conserved sequence element. However, focusing on NS variants may not be overly informative as NS SNPs and synonymous SNPs share similar likelihood and effect size for disease association and synonymous SNPs are just as likely to be involved in disease mechanisms [50]. The remaining variants occurred in the 12S RNA, HV and non-coding regions. Several of the coding and non-coding variants were predicted to belong to evolutionary conserved regions and may play important roles in mtDNA copy number [51] and genome transcription [52, 53] possibly causing severe alterations in mitochondrial function [28, 53-56]. In the present study, mt296 was observed in 38% of haplogroup L cases and no controls, possibly reflecting a risk factor related to mtDNA copy number. Previous studies have described a variant located at mt295 that defines the Caucasian haplogroup J and that has been found to change mitochondrial copy number [51], which may partially account for observations that haplogroup J is over-represented in long-lived people and centenarians from several populations [57-59].

Because collections of rare variants within genes or genomic regions are likely to influence phenotypic expression in important ways [48], examining the collective frequency of rare or singleton variants may reveal the role of specific genes in disease etiology. Our results provide evidence for a significant aggregate effect of sequence variants in the 22 mitochondrial tRNAs for pancreatic cancer risk and suggest that Complex III (*CytB*) gene variants may also play a role. The burden of singleton variants among European haplogoup N cases was three times higher than in controls for the HV2 and 12S RNA regions, suggesting that these mtDNA regions may contribute to risk among persons of European descent. Further, the burden of singleton variants among the African haplogroup L cases was higher than in controls for the entire mtDNA, in particular for OXPHOS complexes I and IV. More specifically, among the complex I genes *ND4* and *ND5*, rare variants were 2-3 times more frequent in cases than in controls whereas in the complex IV *COIII* gene singleton variants occurred in 41% of cases and in no controls. Interestingly, these results are consistent with a study of prostate cancer where germline mtDNA *COI* (complex IV) missense mutations were reported in 11% of prostate cancer cases compared with 2% of the no-cancer controls [28]. This may be of particular importance as incidence and mortality of pancreatic cancer are 48% and 37% higher, respectively, in African-Americans relative to European-Americans [60], and cannot be attributed to racial differences in currently known risk factors [61].

The results of this study suggest that aggregated common and rare variants and the accumulation of singleton variants are important contributors to pancreatic cancer risk. This study had a number of strengths, including: complete mtDNA sequencing allowing for an unbiased assessment of mitochondrial genomic variation; a well-characterized case-control study of pancreatic cancer; and an analytic approach that includes both aggregated and accumulated sequence variants. A few weaknesses are also acknowledged, including: small sample sizes for the African and Asian ancestry samples; low power to detect effects of individual variants; possible survival and selection bias, and no validation study. Demographic characteristics of the study cases who provided blood samples for analyses were similar to those who did not whereas control participants who provided a blood sample were more likely to be white men. While the 13 mtDNA-encoded OXPHOS genes are essential to mitochondrial energy production and are considered the most functionally important [4], hundreds of nuclear DNA-encoded and dozens of mtDNA-encoded bioenergetics genes are distributed throughout both genomes [1, 2]. Future studies of mitochondrial genetic variation will therefore need to account for a complex set of interactions involving the nuclear and mitochondrial genomes [62].

### *2.3.6 Funding sources*

Support provided in part by grants CA133080 (G.J. Tranah, PI), CA59706, CA108370, CA109767, CA89726 (E.A. Holly, PI) from the National Cancer Institute, National Institutes of Health, by the Rombauer Pancreatic Cancer Research Fund and by the Hasbun Fund. The collection of cancer incidence data for the UCSF study was supported by the California Department of Public Health as part of the statewide cancer reporting program; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract N01-PC-35136 to the Northern California Cancer Center; and the Centers for Disease Control and Prevention's National Program of Cancer Registries.

### *2.4 Mitochondrial sequencing in samples from the Health ABC cohort*

A subset of samples from the Heath ABC cohort was chosen for mitochondrial sequencing. I will first describe the Health ABC population in Section 2.4.1. In Sections 2.4.2 and 2.4.3, I will detail our work studying how mitochondrial sequence variation affects energy expenditure and risk of dementia.

#### *2.4.1 Health ABC population*

Participants were part of the Health ABC Study, a prospective cohort study of 3,075 community-dwelling black and white men and women living in Memphis, TN, or Pittsburgh, PA, and aged 70–79 years at recruitment in 1996-1997 [63]. To identify potential participants, a random sample of white and all black Medicare-eligible elders, within designated zip code areas, were contacted. To be eligible, participants had to report no difficulty with activities of daily living, walking a quarter of a mile, or climbing ten steps without resting. They also had to be free of life-threatening cancer diagnoses and have no plans to move out of the study area for at least three years. The sample was approximately balanced for sex (51% women) and 41% of participants were black. Participants self-designated race/ethnicity from a fixed set of options (Asian/Pacific Islander, black/African American, white/Caucasian, Latino/Hispanic, do not know, other). The study was designed to have sufficient numbers of black participants to allow estimates of the relationship of body composition to functional decline. All eligible participants signed a written informed consent, approved by the institutional review boards at the clinical sites. The study was approved by the institutional review boards of the clinical sites and the coordinating center (University of California, San Francisco).

### *2.4.2 Mitochondrial DNA sequence variation and free-living activity energy expenditure in the elderly*

We examined the role of human mitochondrial sequence variation in free-living activity energy expenditure. Several highly conserved and potentially functional variants in OXPHOS genes are unique to samples in the extremes of activity energy expenditure, and there was evidence of variants jointly contributing to the phenotype.

#### *2.4.2.1 Abstract*

The decline in activity energy expenditure underlies a range of age-associated pathological conditions, neuromuscular and neurological impairments, disability, and mortality. We examined the role of mitochondrial genomic variation in free-living activity energy expenditure (AEE) and physical activity levels (PAL) by sequencing the entire mtDNA from 138 Health, Aging, and Body Composition Study participants. Several unique nonsynonymous variants were identified in the extremes of AEE with some occurring at highly conserved sites predicted to affect protein structure and function. Of interest is the p.T194M, *CytB* substitution in the lower extreme of AEE occurring at a residue in the Qi site of complex III. Among participants with low activity levels, the burden of singleton variants was 30% higher across the entire mtDNA and OXPHOS complex I when compared to those having moderate to high activity levels. A significant pooled variant association across the hypervariable 2 (HV2) region was observed for AEE and PAL. These results suggest that mtDNA variation is associated with free-living AEE in older persons and provide an explanation for which specific

mtDNA complexes, genes, and variants may contribute to the maintenance of activity levels in late life.

### *2.4.2.2 Introduction*

Activity energy expenditure (AEE) decreases with age [64, 65] and this decline is associated with an increased risk of mortality, disability, neuromuscular and neurological impairments, and a range of age-associated pathological conditions [66]. Higher free-living AEE is strongly associated with lower risk of mortality among older adults [67]. Manini et al. [67] showed that for every 287 kcal/d in free-living activity energy expenditure (approximately 1 1/4 hours of activity per day), there is approximately a 32% lower risk of mortality. Higher levels of physical activity are associated with reductions in coronary heart disease [68], cancer incidence [69], falls [70], and physical disability [71]. It is unknown, however, why energetic decline occurs and how AEE protects older adults from physical disability, disease and premature mortality. The factors that determine energy balance vary between persons and are to some extent genetically determined [72-77]. The heritability for AEE is 72% [77] and genetic factors explain 30-47% [74, 76] of the variance in resting metabolic rate.

Mitochondrial oxidative phosphorylation (OXPHOS) enzyme activities decline with age in human and primate muscle [78-80], liver [81], and brain [82, 83] and correlate with the accumulation of somatic mitochondrial DNA (mtDNA) deletions [84-109] and base substitutions [110-114]. During the lifetime of an individual, mtDNA undergoes a variety of mutation events and rearrangements that may be important factors in the age-related decline of somatic tissues [115-119]. The progressive and gradual accumulation of mtDNA mutations has been hypothesized to account for the decrease in scope of activity affiliated with the reduced function of cells and organs that accompany the aging process [120]. Impaired mitochondrial function resulting from mtDNA and/or nuclear DNA variation is likely to contribute to an imbalance in cellular energy homeostasis, increase in oxidative stress, and accelerate or inappropriately terminate senescence and aging.

The evolution of human mtDNA is characterized by the emergence of distinct lineages (haplogroups) associated with the major global ethnic groups. It is clear that European ancestry is linked with energy expenditure [121], but in a recent effort we identified specific major African and European haplogroups that had significantly different resting metabolic rate (RMR) and total energy expenditure (TEE) [122]. Both RMR and TEE were significantly elevated in the major European haplogroup N compared to the major African haplogroup L and significant heterogeneity was observed within the African and European lineages [122]. These results demonstrate that mtDNA variants underlying specific haplogroups affect human RMR and TEE and therefore motivate the additional investigation mtDNA sequence-level associations with free-living activity energy expenditure.

While it is clear that AEE levels are associated with environmental factors, mtDNA mutations could have implications for the degree to which physical activity is performed daily. For example, individuals who harbor certain mtDNA mutations would be unable to effectively optimize mitochondria’s ability to rephosphorylate ATP for cellular activities. Following our previous results wherein heterogeneity in TEE was observed among European mitochondrial lineages [122], we sequenced the entire mitochondrial genome in these subjects to examine specific mtDNA variants and aggregate sequence variation that influences differences in AEE.

#### *2.4.2.3 Methods*

We examined the role of mtDNA sequence variation in metabolic rate and energy expenditure by sequencing the entire mtDNA from 138 participants from the Health, Aging, and Body Composition Study. The role of individual variants was first assessed in these phenotypes with an emphasis on nonsynonymous (NS) variants at the extremes of free-living AEE. *In-silico* methods were employed to examine mtDNA nucleotide conservation and predict the functional impact of NS substitutions on amino acid protein sequences. We then examined the collective effects of variants within genes or genomic regions using several rare variant burden tests and assessed singleton burden and substitution rates for evidence of adaptive selection.

*Metabolic rate and energy expenditure measurement*. In 1998-1999, free-living activity energy expenditure was assessed in 302 high-functioning, community-dwelling older adults (aged 70-82 years) from the Health ABC study [67]. The present sequencing study is focused on 138 Health ABC participants of European genetic ancestry with measured free-living AEE. Briefly, RMR was measured via indirect calorimetry on a Deltatrac II respiratory gas analyzer (Datex Ohmeda Inc.); detailed procedures have been described elsewhere [123]. TEE was measured using what is considered the gold-standard and involves a 2-point doubly-labeled water technique that has been previously described [124]. Free-living activity energy expenditure was expressed in two ways [125]. AEE was calculated as [(total energy expenditure*0.90) − resting metabolic rate], removing energy expenditure from the thermic effect of meals that is estimated at 10% of TEE and subtracting energy devoted to basal metabolism. Physical activity level (PAL) is another method for expressing energy expenditure due to physical activity and was calculated as a ratio of TEE and RMR (TEE/RMR). The division of TEE by RMR, a major determinate of which is lean mass, adjusts for differences in body composition (in part reflecting weight and sex) [65]. The PAL formula was adopted by the Food and Agriculture Organization, the World Health Organization, and the United Nations University [126] and these agencies have developed physical activity level categories (sedentary: 1.40-1.69; active, 1.70- 1.99; vigorous activity, 2.00-2.40). AEE and PAL are highly correlated in the current analysis (*r*=0.87) but we provide results for both energy expression types since these offer different advantages (e.g., simplicity of expression and inherent control for differences in body composition, respectively).

*Analysis of individual variants*. Rare sequence variants (minor allele frequency [MAF] <5%) were identified from 48 participants in the extremes (±1SD from the mean) of free-living energy expenditure (AEE < 401 kcal/d vs. 907 kcal/d). These included rare variants from the OXPHOS coding regions (both NS and synonymous [S]), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and each of the three hypervariable (HV) regions. Several *in-silico* methods were employed to examine mtDNA nucleotide conservation (PhastCons [42] and PhyloP [43]) for all variants and to predict the potential functional impact of NS substitutions on amino acid protein sequences (SIFT [127, 128], MutPred [129], and PolyPhen2 [44]). The potential effects of NS substitutions on *CytB*, *COI*, *COII*, and *COIII* were examined with the PyMOL molecular visualization system (v1.4) using the bovine mitochondrial bc1 complex structure with antimycin bound (PDB 2A06, 2.28 Å resolution) [130] and complex IV reduced (PDB 2EIJ, 1.9 Å resolution) and oxidized (PDB 2DYR, 1.8 Å resolution) structures [131, 132].

*Regression.* For individual common mtDNA sequence variants (MAF ≥ 5%) we compared differences in RMR, TEE, AEE and PAL for each allele (mtDNA is single-copy) using a generalized linear model in R (v 2.12.0). All analyses were adjusted for age, sex, lean mass, and 10 eigenvectors of mitochondrial genetic ancestry derived from principal component analysis (PCA, calculated using SAS version 9.1, SAS Institute Inc, Cary, NC).

*Analysis of Aggregated Variants***.** The joint effects of all mitochondrial variants within each gene on energetic measures of interest were evaluated using several rare variant burden tests. Pooled associations of all sequence variants were run using VT test [133] in R and included the T1 (1% MAF threshold) [134], T5 (5% MAF threshold) [134], WE (weighted-sum) [41], and VT (variable threshold) approaches [135]. Energetic measures were adjusted for covariates of age at exam, sex and study site using residuals from linear regression and then normalized to Z scores prior to analyses. We applied these approaches to the four energetic traits and computed statistical significance for each test using 10,000 independent simulations. Variant aggregations were tested across the following regions: 1) the individual OXPHOS complexes; 2) all rRNAs combined; 3) all tRNAs combined; and 4) each of the three HV regions.

*Singletons*. Singletons are variants occurring in single participants that can be quantified to identify genes or genetic regions that harbor significantly higher mutation burdens between groups (e.g. cases vs. controls or phenotype extremes) and possibly play a role in the etiology of a particular disease or trait. Fisher's exact tests were used to compare the total number of singleton variants between participants with little activity (PAL<1.70) and moderate to high activity (PAL≥1.70) for: 1) the entire mitochondrial genome; 2) the individual OXPHOS complexes; 3) the individual genes encoding OXPHOS complexes; 4) all tRNAs combined; 5) all rRNAs combined; and 6) each of the HV regions.

### *2.4.2.4 Results*

A total of 135 Health ABC participants yielded sequence data of sufficient quality for analysis. Of these, 63 were men and 72 were women, with mean (SD) age of 73.4 (2.9) years. Six participants were missing doubly-labeled water measurements resulting in a sample size of 129 for analyses involving TEE, AEE and PAL. Sequencing of 16,544 mtDNA bases (positions 12-16,555) from 135 participants yielded a cumulative total of 449 variants including: 56 common (MAF ≥5%), 160 low frequency variants (MAF 1-5%), and 233 singletons. The 10 duplicate samples had >98% sequence concordance (the majority of discordant calls resulted from positions successfully called in one but called as "N" in another). The within-chip error rate was 0.0028%, which is comparable to previously published rates of 0.0025% and 0.0021% [26, 136].

#### *Individual Variants*

We identified large number of unique OXPHOS, rRNA, tRNA and HV region variants in the extremes of AEE with some occurring at sites that are highly conserved and predicted to affect protein structure or function (Tables 2.4). Most substitutions were unique to single individuals including six *CytB* NS substitutions unique to high and low AEE. Of these, several were predicted to significantly affect function: p.T61A; p.D171N; p.I338V; and p.N374D , and/or to be highly conserved: p.A191T; p.T194M; and p.N374D. Examining the structural model of bovine cytochrome bc1 complex identified the p.A191T, *CytB* and p.T194M, *CytB* substitutions as occurring in a potentially functionally relevant site (Figure 2.4). Some substitutions observed in multiple samples were consistently unique to high ( p.T533M, *ND5*) or low (p.I338V, *CytB*) AEE levels. Two additional variants in the HV2 region were observed in multiple samples that were consistently unique to high (m.200A>G) or low (m.263G>A) AEE levels.

Removing common variants found to be in complete LD ($r^2$=1) yielded 299 “independent” SNPs. Among the 299 “independent” variants there were no individual SNPs that met statistical significance after adjustment for multiple comparisons ($p \leq 1.7*10^{-4}$). MtDNA variants with the strongest effects (in SD units) on AEE and PAL include: m.16324T>C in the HV1 region (AEE=1.9 SD units, p=0.002; PAL=2.1 SD units, p=0.002); m.3609C>T in the *ND1* gene and m.9983A>G in the *COIII* gene (AEE=2.4 SD units, p=0.006); and m.4646T>C in the *ND2* gene (AEE=1.7 SD units, p=0.007; PAL=2.2 SD units, p=0.002).

### ***Pooled Variants***

Significant pooled effects ($p \leq 0.01$ due to multiple test correction) across the HV2 region were observed for free-living AEE and PAL using the T5, WE, and VT methods [135] but not the T1 method (Table 2.5). No statistically significant associations for RMR and TEE were observed for pooled HV2 effects (Table 2.5). Pooled associations for variants across the OXPHOS complexes, rRNAs, and tRNAs were not observed.

A higher burden of singleton variants among sedentary participants was observed across the entire mtDNA (p=0.004), with nominal differences in OXPHOS complex I (p=0.045), *ND4* (p=0.015) and *COI* (p=0.012) when compared with active participants (Figure 2.5). The frequency of singletons across the entire mtDNA and complex I was 30% higher in sedentary versus active participants. The frequency of singleton variants in the *ND4* and *COI* genes was 2-3 times higher in sedentary versus active participants. By contrast, the proportion of singleton variants in the *ND4L* (p=0.03) and *COII* genes (p=0.03) was 10 times higher in the active group when compared with the sedentary group (Figure 2.5).

### *2.4.2.5 Discussion*

We examined the role of mtDNA sequence variation in AEE and identified a large number of highly conserved and potentially functional variants that are unique to individuals with either extremely high or low AEE. Among these are variants that have been implicated in mitochondrial several diseases, including: Leber's Hereditary Optic Neuropathy (LHON); mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS); and mitochondrial cardiomyopathy. We also considered how multiple sequence variants influence energy levels and identified aggregated variation and accumulations of singletons that also impact energy expenditure.

Among the six *CytB* NS substitutions unique to high and low AEE levels, several were predicted to significantly affect function and/or to be highly conserved. Of particular interest are the p.A191T, *CytB* and p.T194M, *CytB* substitutions which are unique to participants with low AEE. Both are located in the Qi binding pocket of complex III, where quinone is reduced by cytochrome b [137]. The p.T194M, *CytB* variant occurs at a residue that is noted to undergo significant conformational changes upon contact with antimycin A, a pharmacological inhibitor of the Qi site [137]. In the presence of antimycin A, complex III produces high quantities of superoxide indicating that inhibition at this site blocks electron transfer (from cytochrome b to quinone at Qi) causing a buildup of semiquinone at the Qo site. This buildup results in increased ROS production from complex III [6]. The structure of bovine cytochrome bc1 complex was also used to predict whether specific *CytB* NS substitutions occur in functionally relevant sites. The p.N374D, *CytB* substitution occurs near Lysine (311, 375, and 378) and Serine (310, 314, and 370) residues and may be potentially involved in polar interactions with these neighboring sites. The p.D171N, *CytB* substitution which is located on the outer core of protein is a risk factor for LHON [138-145]. Complex III is the ETC enzyme responsible for oxidizing ubiquionol and transferring electrons to cytochrome c through the cytochrome b mediated Q cycle. During the process of electron transfer through complex III, a net of 4 protons are pumped out of the mitochondrial matrix increasing PMF. The resulting reduced cytochrome c then transports the electrons downstream to complex IV. If the mutations identified by sequencing lead to dysfunction in cytochrome b, the result may be a backup of electrons in the upstream OXPHOS components resulting in ROS production and insufficient ATP supply [8].

Among the complex I NS substitutions identified in the extremes of AEE, several are predicted to impact function including two that are considered possible risk factors for LHON: p.I57M, *ND2* [146] and p.Y159H, *ND5* [140, 147]. The p.I57M, *ND2* substitution is predicted to cause a gain of a catalytic residue and a gain of disorder. In addition, the p.M1T, *ND1* substitution is a risk factor for MELAS [148] and the m.3308T>C mutation that encodes this substitution may alter the hydrophobicity and antigenicity of the N-terminal peptide of *ND1* [148]. Other substitutions are predicted to results in the loss of stability p.I96T, *ND3*, loss of a catalytic residue p.T533M, *ND5*, and the gain of a catalytic residue p.I100V, *ND5*.Complex I is a large multi-subunit, membrane-bound protein which serves as the major entry point for most electrons into the electron transport chain (ETC). This process involves the electron transfer from NADH to quinone and contributes to the generation of mitochondrial proton motive force (PMF, potential energy for ATP generation) through the pumping of 4 protons. In eukaryotes, the mitochondrial genome encodes the 7 most hydrophobic subunits of complex I (*ND1-ND6* and *ND4L*) [149, 150]. These proteins comprise a large portion of the membrane domain in complex I and are thought to be essential to both quinone binding and proton translocation.

Among the complex V NS substitutions identified in the extremes of AEE are p.P10S, *ATP8* and p.M42T, *ATP8*.Complex V is a multisubunit complex consisting of two functional domains, $F_1$ and $F_0$. The $F_0$ domain is embedded in the mitochondrial inner membrane and is in part encoded by the mtDNA *ATP6* and *ATP8* genes. Complex

V is the site of ATP synthesis, a process that consumes membrane potential by allowing protons to flow back down their electro-chemical gradient into the mitochondrial matrix, resulting in ATP production. Defects in complex V are associated with ATP synthase deficiency and it has been proposed that mutations in *ATP6* and *ATP8* are associated with reduced complex V assembly and impaired ATP synthase function [151, 152]. Potentially, modification of the function of these integral components of the ETC could alter the efficiency of ATP production or result superoxide production through a backup of electrons on the upstream ETC components.

Two variants in the HV2 region were observed in multiple samples that were consistently unique to high (m.200A>G) or low (m.263G>A) AEE levels. While it is not clear how these HV2 variants impact AEE, it is possible that this variation is involved in regulating mtDNA copy number [51]. The functions of the HV2 region include: priming site for mtDNA replication; the heavy-strand origin encoding 12 of the 13 OXPHOs genes; three conserved sequence blocks; and two transcription factor binding sites [153]. In a previous study the HV2 m.295C>T variant was found to increase both mtDNA transcription and copy number [51]. This particular mtDNA variant defines Caucasian haplogroup J and cybrids (experimental hybrid cells containing mtDNA from different sources placed in a uniform nuclear DNA background) containing haplogroup J mtDNA had a greater than 2-fold increase in mtDNA copy number compared with cybrids containing haplogroup H mtDNA [51]. The impact of the haplogroup J regulatory region mutation on mtDNA replication or stability may partially account for several observations that haplogroup J is over-represented in long-lived people and centenarians from several populations [57-59]. Several variants in the tRNA and rRNA regions were observed in samples that were consistently unique to high or low AEE levels. The mitochondrial tRNAs and rRNAs are critical for protein synthesis and mitochondrial assembly. The m.8348A>G (tRNA Lys) variant that is unique to a participant with extremely low AEE has also been identified as a risk factor for cardiomyopathy [154].

As collections of variants within genes or genomic regions are likely to influence phenotypes in important ways [48], examining the combined effect of rare variants may also reveal the role of specific genes in disease etiology. Across the entire mtDNA and complex I specifically we observed a significant 30% higher singleton burden among sedentary participants when compared to those defined as active. In addition, the singleton burden for *ND4* and *COI* genes was twice as high in sedentary participants whereas the proportion of singleton variants in the *ND4L* and COII genes was 10 times higher in the active group. Complex I is a major contributor to cellular reactive oxygen species (ROS) production [155].  Inhibition of complex I leads to increased generation of ROS, decreased ATP levels, and induction of apoptosis [7, 156, 157], all of which could play a major role in reducing AEE. Dysfunction in complex I has been linked to multiple diseases and mitochondrial pathologies including tumorigenesis [158], Parkinson's disease [159], and aging [160] (through a ROS dependent or a ROS independent mechanism). The only region exhibiting evidence of positive selection was complex III, suggesting that *CytB* may be undergoing adaptive selection.

Analytic approaches that test the combined effect of multiple variants have been used to resolve genetic associations for several complex traits [161-164] including the role of rare mitochondrial variants in disease [165]. We evaluated several approaches including the allele-frequency threshold approach (1% or 5%) [134], a weighted-sum approach [41], and the variable-threshold approach [135]. Significant variant burden effects in the HV2 region were observed for free-living energy expenditure. Rare variant burden in HV2 was associated with AEE and PAL but not with RMR or TEE, suggesting that this variation is most important for physical activity and volitional exercise [166]. Our results also suggest that HV2 variation under the 5% allele-frequency threshold, but not under the 1% allele-frequency threshold is associated with AEE and PAL, though this finding may be due to a lack of statistical power. Both weighted-sum and variable-threshold approaches, however, suggest that HV2 variation is associated with AEE and PAL.

This study had a number of strengths, including: complete mtDNA sequencing allowing for an unbiased assessment of mitochondrial genomic variation; a well-characterized population-based longitudinal cohort with energetics measured using state of the art methods; an analytic approach that includes both aggregated and accumulated sequence variants; and *in silico* prediction and structural modeling that allowed for detailed interpretation of sequence-based findings. Some weaknesses are also acknowledged, including: small sample size and low power to detect an effect of individual variants. It is possible that the mtDNA variants identified in this study may not be causally related to the energetic phenotypes thus the lack of a replication cohort is also a limitation.

In summary, there is little understanding of genetic factors that contribute to an individual's daily activity levels and here we identify a number of potentially functional mtDNA variants and collections of sequence variants that contribute to free-living activity energy expenditure. These results may help to uncover specific mitochondrial functions that explain age-related declines in activity but also maintenance of high activity energy levels in the elders. While the 13 mtDNA-encoded OXPHOS genes are essential to mitochondrial energy production and are considered the most functionally important [4], hundreds of nuclear DNA-encoded and dozens of mtDNA-encoded bioenergetics genes are distributed throughout both genomes [1, 2]. We have shown that nuclear genomic European Ancestry in African Americans is strongly associated with higher RMR [167]. Future studies of mitochondrial genetic variation will therefore need to account for a complex set of interactions involving the nuclear and mitochondrial genomes [62]. Since the 13 mtDNA-encoded OXPHOS genes are essential to mitochondrial energy production [4] , the coding region variation identified in this study might be related to ROS production at OXPHOS complexes I and III, ATP generation efficiency through the collective impairment of the respiratory chain [5, 6] or through apoptosis [7]. Individual and collective variation in the HV2, tRNA and rRNA regions may affect mitochondrial function by impacting the rate or efficiency of mitochondrial biogenesis (increase in mitochondrial number and/or mass). An important aspect of mitochondrial biogenesis is rate of turnover, which is thought to decline with age [168]. Impaired ability to turnover may allow for defective mitochondria to accumulate, especially in older, postmitotic cells lead to impaired respiratory capacity [169]. It is known that mitochondrial biogenesis is affected by pharmacologic agents [170-175], natural compounds such as resveratrol [176] and behavioral interventions such as caloric restriction and exercise [177-180]. However, by identifying mitochondrial genetic variants that influence free-living activity energy expenditure we provide evidence that additional molecular targets (*e.g.* Qi binding pocket of complex III) or mechanisms (*e.g.* mitochondrial protein synthesis and assembly) could be involved.

#### *2.4.2.6 Funding sources*

This research was supported by National Institute on Aging (NIA) Contracts N01-AG-6-2101; N01-AG-6-2103; N01-AG-6-2106; NIA grants R01-AG028050 and R03-AG032498, NINR grant R01-NR012459; and Z01A6000932. E.T.L. was supported in part by NIH Training Grant T32 GM007175 and Y.Z. by NLM grant LM009722. Data analyses for this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (http://biowulf.nih.gov).

### *2.4.3 Mitochondrial DNA sequence variation and dementia and cognitive function in the elderly*

Driven by the observation of mitochondrial dysfunction in patients with Alzheimer's disease, this study examines how mitochondrial sequence variation contributes to cognitive decline and dementia risk in a subset of samples form the Health ABC study.

#### *2.4.3.1 Abstract*

Mitochondrial dysfunction is a prominent hallmark of Alzheimer's disease (AD), and mitochondrial DNA (mtDNA) damage may be a major cause of abnormal ROS production in AD. The purpose of this study was to assess the influence of mtDNA sequence variation on clinically significant cognitive impairment and dementia risk in the

population-based Health, Aging, and Body Composition (Health ABC) Study. All participants were free of dementia at baseline and incidence was determined from hospital and medication records over 10-12 years of follow-up. The Modified Mini-Mental State Examination (3MS) and Digit Symbol Substitution Test (DSST) were administered at baseline. We sequenced the complete ~16.5kb mtDNA from 135 Health ABC participants and identified several highly conserved and potentially functional nonsynonymous variants unique to 22 dementia cases and aggregate sequence variation across the hypervariable 2-3 regions that influences 3MS and DSST scores.

#### *2.4.3.2 Introduction*

The prevalence of cognitive decline and dementia, largely in the form of Alzheimer's disease (AD), affects approximately 10% of adults over the age of 65 rising exponentially to 50% of adults over the age of 85 in the United States [181]. Several conserved mechanisms underlie the changes observed in the aging brain including mitochondrial function and oxidative stress, autophagy and protein turnover [182]. Considerable evidence suggests that the changes in mitochondria and oxidative stress levels precede plaques, tangles and clinical manifestation of AD in humans [183]. Changes in mitochondrial function are causally linked to several early abnormalities that accompany AD including plaques and tangles [183]. Hence, the early alterations to mitochondria, which can induce multiple abnormalities, may present more desirable therapeutic targets than the reversal of the individual pathologies that occur in later cognitive decline and dementia.

Increased oxidative damage [184] and mitochondrial dysfunction are important early factors for accelerated cognitive decline and AD [185, 186]. Measures of ROS damage to proteins, nucleic acids, carbohydrates and lipids are found in cerebrospinal fluid, plasma, and urine of subjects at very early stages of AD [187] and in autopsy brains from AD patients [188-192]. Different forms of mitochondrial impairment or oxidative stress can mimic AD-like changes to the brain [193-203]. For example, induction of oxidative stress by multiple approaches increases amyloid-ß (Aß) production and plaque formation [204-208]. In fact, oxidative damage in AD brain is more pervasive than plaques and tangles [209] and changes in mitochondrial function have been shown to precede Aβ and tangle formation [210-215]. Mitochondria-derived ROS [216-218] and Aß formation [216, 219] may create a cycle that further exacerbates mitochondrial dysfunction [169] and accelerates cognitive impairment [219]. Organ-specific analysis of brain aging has revealed a progressive decline in mitochondrial gene expression in rats, rhesus macaques and humans [194, 220, 221]. Several lines of evidence show that key enzymes responsible for mitochondrial energy metabolism are severely affected in AD [222-224] with some genes coding for respiratory chain subunits being differentially expressed in AD patients [225]. In particular, oxidative phosphorylation (OXPHOS) defects resulting from somatic mtDNA mutations may play a role in AD pathophysiology [226]. The brain is particularly susceptible to defective mitochondrial function related to mitochondrial DNA (mtDNA) mutations [182, 183].

Individual mtDNA mutations have been identified in patients with AD [224, 227-241]; however, these studies were small and most of the identified variants have not been confirmed [234, 236-241]. In the present study, we sequenced the complete ~16.5kb mtDNA from 135 Health ABC participants to identify both highly conserved and potentially functional mutations unique to dementia cases and aggregate sequence variation that influences tests of cognitive function.

### *2.4.3.3 Methods*

*Cognitive function testing*. The Modified Mini-Mental State Examination (3MS) was administered to participants at baseline (year 1) and after two, four, and seven years of follow-up (years 3, 5, and 8). The 3MS is a brief, general cognitive battery with components for orientation, concentration, language, praxis, and immediate and delayed memory [242]. Possible scores range from 0 to 100, with higher scores indicating better cognitive function. The Digit Symbol Substitution Test (DSST) was administered at years 1, 5, and 8. The DSST measures response speed, sustained attention, visual spatial skills, and set shifting, all of which reflect executive cognitive function [243, 244] . The test is reported to distinguish mild dementia from healthy aging [245]. The DSST score is calculated as the total number of items correctly coded in 90 seconds, with a higher score indicating better cognitive function. Participant-specific slopes of DSST scores were estimated from mixed-effects models with random intercepts and slopes [246]. The participant-specific slopes of 3MS scores were estimated by best linear unbiased predictions using a linear mixed model with random intercepts and slopes.

*Dementia incidence*. All participants were free of dementia at baseline. Incident dementia was determined by the date of the first available record of a dementia diagnosis over 10-12 years of follow-up. Cases were identified through hospital records indicating a dementia related hospital event, either as the primary or secondary diagnosis related to the hospitalization, or by record of prescribed dementia medication (i.e. galantamine, rivastigmine, memantine, donepezil, tacrine).

*Variant analysis*. Rare sequence variants and singletons unique to dementia cases were identified from the OXPHOS coding regions. Several *in-silico* methods were employed to examine mtDNA nucleotide conservation (PhastCons [42] and PhyloP [43]) for all mtDNA variants and to predict the potential functional impact of NS substitutions on amino acid protein sequences (SIFT [127, 128], MutPred [129], and PolyPhen2 [44]). The joint effects of all mitochondrial variants within each gene or region on cognitive function test scores were evaluated using VT test [133] in R. All analyses were adjusted for age at baseline, sex, and study site using residuals from linear regression and then normalized to Z scores prior to analyses. We applied these approaches to baseline 3MS and DSST scores and computed statistical significance for each test using 10,000 independent simulations. Variant aggregations were tested across the following regions: 1) the individual OXPHOS complexes; 2) all rRNAs combined; 3) all tRNAs combined; and 4) each of the HV regions.

### *2.4.3.4 Results*

Among the subset of 135 sequenced participants, 22 (16%) developed dementia (Table 2.6). Dementia cases were more likely to have had diabetes and be APOEe4 allele carriers and were also more likely to be female and less likely to have a postsecondary education (Table 2.6).

Sequencing of 16,544 mtDNA bases (positions 12-16,555) from 135 participants yielded a cumulative total of 449 variants including: 56 common (MAF ≥5%), 160 low frequency variants (MAF 1-5%), and 233 singletons. Among 22 participants with incident dementia we identified 10 NS variants with some occurring at highly conserved sites predicted to affect protein structure/function (Table 2.7). MutPred predicted a gain of acetylation for the *ATP8* E52K substitution (p=0.004). The *CytB* p.A191T and p.T194M substitutions occur in a potentially functionally relevant site known as the Qi binding pocket. Among non-coding variants unique to dementia cases were two HV2 (m.114, C>T; m.238, A>T), two 16S rRNA (m.1700, T>C; 1700t/c, m.2141, T>C), and three tRNA (m.5527, A>G; m.5567, T>C; m.5592, A>G) variants. Nominally significant pooled effects across HV2 were observed for 3MS (p=0.04, T1 method) and HV3 for DSST (p=0.04, VT method); however, these tests were not significant after multiple test correction for 8 mtDNA regions (critical a=0.006).

### *2.4.3.5 Discussion*

In this study, we sequenced the entire mtDNA in a large, population-based, longitudinal study of elderly participants to examine the role of rare sequence variants, and aggregate mitochondrial sequence variation in determining dementia risk and cognitive decline. We identified several highly conserved and potentially functional variants that were unique to dementia cases. Of particular interest are the *CytB* p.A191T and p.T194M substitutions located in the complex III Qi binding pocket, where quinone is reduced by cytochrome b [137]. The p.T194M, *CytB* variant occurs at a residue that is noted to undergo significant conformational changes upon contact with antimycin A, a pharmacological inhibitor of the Qi site [137].

The gain of lysine p.E52K, *ATP8* can result in a new target for a diverse array of post-translational modifications which can impact the protein function, including acetylation [247]. This substitution may facilitate acetylation-directed molecular interactions and in this case impact ATP8 activity and stability [248-250]. Two additional dementia-related variants include p.Y159H, *ND5* which is a possible risk factors for Leber's Hereditary Optic Neuropathy [140, 147], and p.A331D, *ND2* which was previously been associated with Alzheimer’s Disease [227] although the role of this variant remains unclear [237, 238]. Several variants in the tRNA, rRNA, and HV2 regions were unique to dementia cases. The mitochondrial tRNAs and rRNAs are critical for protein synthesis and mitochondrial assembly and the HV2 region includes the priming site for mtDNA replication and the heavy-strand origin encoding 12 of the 13 OXPHOS genes [153].

We also considered how multiple sequence variants influence cognitive function and identified aggregated variation impacting the 3MS and DSST. Variant pooling effects suggest rare HV2 and HV3 sequence variation is associated with 3MS and DSST, respectively. It is not clear how HV2 and HV3 sequence variation impact cognitive function; however, HV2 and HV3 are involved in regulating mtDNA copy number [51, 153]. Mitochondrial DNA copy number in peripheral blood has been inversely correlated with insoluble Aß40 and Aß42 levels [185] and is positively correlated with cognitive function in healthy elderly women [251].

Individual and collective variation in the HV2, HV3, tRNA and rRNA regions may affect mitochondrial function by impacting the rate or efficiency of mitochondrial biogenesis (increase in mitochondrial number and/or mass). An important aspect of mitochondrial biogenesis is turnover rate, which is thought to decline with age [168]. Impaired ability to turnover may allow the accumulation of defective mitochondria resulting from oxidative injury [217, 218], Aß accumulation [216, 219], and/or mtDNA damage [252] in neurons leading to impaired respiratory capacity [169]. By identifying mitochondrial genetic variants that influence cognitive function and dementia risk we provide evidence that additional molecular mechanisms (e.g. mitochondrial protein synthesis and assembly) or targets (e.g. Qi binding pocket of complex III) could be involved. Such findings might lead to the development of interventions or new clinical strategies for improving mitochondrial function and delaying the onset of cognitive decline.

### **2.4.3.6** ***Funding sources***

This research was supported by National Institute on Aging (NIA) Contracts N01-AG-6-2101; N01-AG-6-2103; N01-AG-6-2106; NIA grants R01-AG028050 and R03-AG032498, NINR grant R01-NR012459; and Z01AG000951. E.T.L. was supported in part by NIH Training Grant T32 GM007175 and Y.Z. by NLM grant LM009722. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. Data analyses for this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (http://biowulf.nih.gov).

### *2.5 References*

1. Wallace, D.C., *Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine.* Annu Rev Biochem, 2007. **76**: p. 781-821.
2. Wallace, D.C., W. Fan, and V. Procaccio, *Mitochondrial energetics and therapeutics.* Annu Rev Pathol, 2010. **5**: p. 297-348.
3. Giles, R.E., et al., *Maternal inheritance of human mitochondrial DNA.* Proc Natl Acad Sci U S A, 1980. **77**(11): p. 6715-9.
4. Wallace, D.C., *Colloquium paper: bioenergetics, the origins of complexity, and the ascent of man.* Proc Natl Acad Sci U S A, 2010. **107 Suppl 2**: p. 8947-53.
5. Niemi, A.K., et al., *A combination of three common inherited mitochondrial DNA polymorphisms promotes longevity in Finnish and Japanese subjects.* Eur J Hum Genet, 2005. **13**(2): p. 166-70.
6. Brand, M.D., *The sites and topology of mitochondrial superoxide production.* Exp Gerontol. **45**(7-8): p. 466-72.
7. Li, N., et al., *Mitochondrial complex I inhibitor rotenone induces apoptosis through enhancing mitochondrial reactive oxygen species production.* J Biol Chem, 2003. **278**(10): p. 8516-25.
8. Tarnopolsky, M.A., et al., *Attenuation of free radical production and paracrystalline inclusions by creatine supplementation in a patient with a novel cytochrome b mutation.* Muscle Nerve, 2004. **29**(4): p. 537-47.
9. Wallace, D.C., et al., *Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations.* Proc Natl Acad Sci U S A, 1997. **94**(26): p. 14900-5.
10. Neckelmann, N., et al., *cDNA sequence of a human skeletal muscle ADP/ATP translocator: lack of a leader peptide, divergence from a fibroblast translocator cDNA, and coevolution with mitochondrial DNA genes.* Proc Natl Acad Sci U S A, 1987. **84**(21): p. 7580-4.
11. Merriwether, D.A., et al., *The structure of human mitochondrial DNA variation.* J Mol Evol, 1991. **33**(6): p. 543-55.
12. Lam, E.T., et al., *Mitochondrial DNA sequence variation and risk of pancreatic cancer.* Cancer Res, 2011.
13. Symons, S., et al. *ResqMi - a Versatile Algorithm and Software for Resequencing Microarrays*. in *Proceedings of the German Conference on Bioinformatics*. 2008. Dresden, Germany.
14. *American Cancer Society Facts and Figures 2010.* Atlanta, GA: American Cancer Society, 2010.
15. Lowenfels, A.B. and P. Maisonneuve, *Epidemiology and prevention of pancreatic cancer.* Jpn J Clin Oncol, 2004. **34**(5): p. 238-44.
16. Warburg, O.H., *The metabolism of tumours. R.R. Smith, New York, New York.* 1931.
17. Hruszkewycz, A.M. and D.S. Bergtold, *Oxygen radicals, lipid peroxidation and DNA damage in mitochondria.* Basic Life Sci, 1988. **49**: p. 449-56.
18. Hruszkewycz, A.M., *Lipid peroxidation and mtDNA degeneration. A hypothesis.* Mutat Res, 1992. **275**(3-6): p. 243-8.

19. Richter, C., J.W. Park, and B.N. Ames, *Normal oxidative damage to mitochondrial and nuclear DNA is extensive.* Proc Natl Acad Sci U S A, 1988. **85**(17): p. 6465-7.
20. Kroemer, G., *Mitochondria in cancer.* Oncogene, 2006. **25**(34): p. 4630-2.
21. Cuezva, J.M., et al., *The bioenergetic signature of cancer: a marker of tumor progression.* Cancer Res, 2002. **62**(22): p. 6674-81.
22. Carew, J.S. and P. Huang, *Mitochondrial defects in cancer.* Mol Cancer, 2002. **1**: p. 9.
23. Rossignol, R., et al., *Energy substrate modulates mitochondrial structure and oxidative capacity in cancer cells.* Cancer Res, 2004. **64**(3): p. 985-93.
24. Jones, J.B., et al., *Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations.* Cancer Res, 2001. **61**(4): p. 1299-304.
25. Kassauei, K., et al., *Mitochondrial DNA mutations in pancreatic cancer.* Int J Gastrointest Cancer, 2006. **37**(2-3): p. 57-64.
26. Maitra, A., et al., *The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection.* Genome Res, 2004. **14**(5): p. 812-9.
27. Booker, L.M., et al., *North American white mitochondrial haplogroups in prostate and renal cancer.* J Urol, 2006. **175**(2): p. 468-72; discussion 472-3.
28. Petros, J.A., et al., *mtDNA mutations increase tumorigenicity in prostate cancer.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 719-24.
29. Bai, R.K., et al., *Mitochondrial genetic background modifies breast cancer risk.* Cancer Res, 2007. **67**(10): p. 4687-94.
30. Mosquera-Miguel, A., et al., *Is mitochondrial DNA variation associated with sporadic breast cancer risk?* Cancer Res, 2008. **68**(2): p. 623-5; author reply 624.
31. Fang, H., et al., *Cancer type-specific modulation of mitochondrial haplogroups in breast, colorectal and thyroid cancer.* BMC Cancer, 2010. **10**: p. 421.
32. Liu, V.W., et al., *Mitochondrial DNA variant 16189T>C is associated with susceptibility to endometrial cancer.* Hum Mutat, 2003. **22**(2): p. 173-4.
33. Darvishi, K., et al., *Mitochondrial DNA G10398A polymorphism imparts maternal Haplogroup N a risk for breast and esophageal cancer.* Cancer Lett, 2007. **249**(2): p. 249-55.
34. Canter, J.A., et al., *Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women.* Cancer Res, 2005. **65**(17): p. 8028-33.
35. Setiawan, V.W., et al., *Mitochondrial DNA G10398A variant is not associated with breast cancer in African-American women.* Cancer Genet Cytogenet, 2008. **181**(1): p. 16-9.
36. Halfdanarson, T.R., et al., *Mitochondrial genetic polymorphisms do not predict survival in patients with pancreatic cancer.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(9): p. 2512-3.
37. Wang, L., et al., *Mitochondrial genetic polymorphisms and pancreatic cancer risk.* Cancer Epidemiol Biomarkers Prev, 2007. **16**(7): p. 1455-9.
38. Holly, E.A., C.A. Eberle, and P.M. Bracci, *Prior history of allergies and pancreatic cancer in the San Francisco Bay area.* Am J Epidemiol, 2003. **158**(5): p. 432-41.

39. van Oven, M. and M. Kayser, *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.* Hum Mutat, 2009. **30**(2): p. E386-94.

40. Biffi, A., et al., *Principal-component analysis for assessment of population stratification in mitochondrial medical genetics.* Am J Hum Genet, 2010. **86**(6): p. 904-17.

41. Madsen, B.E. and S.R. Browning, *A groupwise association test for rare mutations using a weighted sum statistic.* PLoS Genet, 2009. **5**(2): p. e1000384.

42. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

43. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.

44. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

45. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.

46. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.

47. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.

48. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* Curr Opin Genet Dev, 2009. **19**(3): p. 212-9.

49. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.

50. Chen, R., et al., *Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association.* PLoS ONE. **5**(10): p. e13574.

51. Suissa, S., et al., *Ancient mtDNA genetic variants modulate mtDNA transcription and replication.* PLoS Genet, 2009. **5**(5): p. e1000474.

52. Fernandez-Silva, P., J.A. Enriquez, and J. Montoya, *Replication and transcription of mammalian mitochondrial DNA.* Exp Physiol, 2003. **88**(1): p. 41-56.

53. Penta, J.S., et al., *Mitochondrial DNA in human malignancy.* Mutat Res, 2001. **488**(2): p. 119-33.

54. Hochhauser, D., *Relevance of mitochondrial DNA in cancer.* Lancet, 2000. **356**(9225): p. 181-2.

55. Kang, D. and N. Hamasaki, *Alterations of mitochondrial DNA in common diseases and disease states: aging, neurodegeneration, heart failure, diabetes, and cancer.* Curr Med Chem, 2005. **12**(4): p. 429-41.

56. Yoneyama, H., et al., *Nucleotide sequence variation is frequent in the mitochondrial DNA displacement loop region of individual human tumor cells.* Mol Cancer Res, 2005. **3**(1): p. 14-20.

57. Niemi, A.K., et al., *Mitochondrial DNA polymorphisms associated with longevity in a Finnish population.* Hum Genet, 2003. **112**(1): p. 29-33.

58. Ross, O.A., et al., *Mitochondrial DNA polymorphism: its role in longevity of the Irish population.* Exp Gerontol, 2001. **36**(7): p. 1161-78.

59. De Benedictis, G., et al., *Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans.* Faseb J, 1999. **13**(12): p. 1532-6.

60. http://seer.cancer.gov/csr/1975_2005.
61. Arnold, L.D., et al., *Are racial disparities in pancreatic cancer explained by smoking and overweight/obesity?* Cancer Epidemiol Biomarkers Prev, 2009. **18**(9): p. 2397-405.
62. Tranah, G.J., *Mitochondrial-nuclear epistasis: Implications for human aging and longevity.* Ageing Res Rev, 2011. **10**(2): p. 238-52.
63. Rooks, R.N., et al., *The association of race and socioeconomic status with cardiovascular disease indicators among older adults in the health, aging, and body composition study.* J Gerontol B Psychol Sci Soc Sci, 2002. **57**(4): p. S247-56.
64. Elia, M., P. Ritz, and R.J. Stubbs, *Total energy expenditure in the elderly.* Eur J Clin Nutr, 2000. **54 Suppl 3**: p. S92-103.
65. Black, A.E., et al., *Human energy expenditure in affluent societies: an analysis of 574 doubly-labelled water measurements.* Eur J Clin Nutr, 1996. **50**(2): p. 72-92.
66. Linnane, A.W., *Mitochondria and aging: the universality of bioenergetic disease.* Aging (Milano), 1992. **4**(4): p. 267-71.
67. Manini, T.M., et al., *Daily activity energy expenditure and mortality among older adults.* Jama, 2006. **296**(2): p. 171-9.
68. Wannamethee, S.G., A.G. Shaper, and M. Walker, *Changes in physical activity, mortality, and incidence of coronary heart disease in older men.* Lancet, 1998. **351**(9116): p. 1603-8.
69. Gregg, E.W., et al., *Relationship of changes in physical activity and mortality among older women.* Jama, 2003. **289**(18): p. 2379-86.
70. Gregg, E.W., M.A. Pereira, and C.J. Caspersen, *Physical activity, falls, and fractures among older adults: a review of the epidemiologic evidence.* J Am Geriatr Soc, 2000. **48**(8): p. 883-93.
71. Ferrucci, L., et al., *Smoking, physical activity, and active life expectancy.* Am J Epidemiol, 1999. **149**(7): p. 645-53.
72. Bouchard, C., et al., *Overfeeding in identical twins: 5-year postoverfeeding results.* Metabolism, 1996. **45**(8): p. 1042-50.
73. Bouchard, C., et al., *The response to exercise with constant energy intake in identical twins.* Obes Res, 1994. **2**(5): p. 400-10.
74. Jacobson, P., et al., *Resting metabolic rate and respiratory quotient: results from a genome-wide scan in the Quebec Family Study.* Am J Clin Nutr, 2006. **84**(6): p. 1527-33.
75. Norman, R.A., et al., *Autosomal genomic scan for loci linked to obesity and energy metabolism in Pima Indians.* Am J Hum Genet, 1998. **62**(3): p. 659-68.
76. Wu, X., et al., *A genome scan among Nigerians linking resting energy expenditure to chromosome 16.* Obes Res, 2004. **12**(4): p. 577-81.
77. Joosen, A.M., et al., *Genetic analysis of physical activity in twins.* Am J Clin Nutr, 2005. **82**(6): p. 1253-9.
78. Cooper, J.M., V.M. Mann, and A.H. Schapira, *Analyses of mitochondrial respiratory chain function and mitochondrial DNA deletion in human skeletal muscle: effect of ageing.* J Neurol Sci, 1992. **113**(1): p. 91-8.
79. Boffoli, D., et al., *Decline with age of the respiratory chain activity in human skeletal muscle.* Biochim Biophys Acta, 1994. **1226**(1): p. 73-82.

80. Trounce, I., E. Byrne, and S. Marzuki, *Decline in skeletal muscle mitochondrial respiratory chain function: possible factor in ageing.* Lancet, 1989. **1**(8639): p. 637-9.

81. Yen, T.C., et al., *Liver mitochondrial respiratory functions decline with age.* Biochem Biophys Res Commun, 1989. **165**(3): p. 944-1003.

82. Bowling, A.C., et al., *Age-dependent impairment of mitochondrial function in primate brain.* J Neurochem, 1993. **60**(5): p. 1964-7.

83. Jazin, E.E., et al., *Human brain contains high levels of heteroplasmy in the noncoding regions of mitochondrial DNA.* Proc Natl Acad Sci U S A, 1996. **93**(22): p. 12382-7.

84. Corral-Debrinski, M., et al., *Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age.* Nat Genet, 1992. **2**(4): p. 324-9.

85. Arnheim, N. and G. Cortopassi, *Deleterious mitochondrial DNA mutations accumulate in aging human tissues.* Mutat Res, 1992. **275**(3-6): p. 157-67.

86. Corral-Debrinski, M., et al., *Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease.* Mutat Res, 1992. **275**(3-6): p. 169-80.

87. Wallace, D.C., et al., *Mitochondrial DNA mutations in human degenerative diseases and aging.* Biochim Biophys Acta, 1995. **1271**(1): p. 141-51.

88. Cortopassi, G.A., et al., *A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues.* Proc Natl Acad Sci U S A, 1992. **89**(16): p. 7370-4.

89. Hattori, K., et al., *Age-dependent increase in deleted mitochondrial DNA in the human heart: possible contributory factor to presbycardia.* Am Heart J, 1991. **121**(6 Pt 1): p. 1735-42.

90. Hayakawa, M., et al., *Age-associated damage in mitochondrial DNA in human hearts.* Mol Cell Biochem, 1993. **119**(1-2): p. 95-103.

91. Linnane, A.W., et al., *Mitochondrial gene mutation: the ageing process and degenerative diseases.* Biochem Int, 1990. **22**(6): p. 1067-76.

92. Chang, M.C., et al., *Accumulation of mitochondrial DNA with 4977-bp deletion in knee cartilage--an association with idiopathic osteoarthritis.* Osteoarthritis Cartilage, 2005. **13**(11): p. 1004-11.

93. Yang, J.H., et al., *A specific 4977-bp deletion of mitochondrial DNA in human ageing skin.* Arch Dermatol Res, 1994. **286**(7): p. 386-90.

94. Mann, V.M., J.M. Cooper, and A.H. Schapira, *Quantitation of a mitochondrial DNA deletion in Parkinson's disease.* FEBS Lett, 1992. **299**(3): p. 218-22.

95. Melov, S., et al., *Marked increase in the number and variety of mitochondrial DNA rearrangements in aging human skeletal muscle.* Nucleic Acids Res, 1995. **23**(20): p. 4122-6.

96. Nagley, P., et al., *Mitochondrial DNA mutation associated with aging and degenerative disease.* Ann N Y Acad Sci, 1992. **673**: p. 92-102.

97. Piko, L., A.J. Hougham, and K.J. Bulpitt, *Studies of sequence heterogeneity of mitochondrial DNA from rat and mouse tissues: evidence for an increased frequency of deletions/additions with aging.* Mech Ageing Dev, 1988. **43**(3): p. 279-93.

98. Simonetti, S., et al., *Accumulation of deletions in human mitochondrial DNA during normal aging: analysis by quantitative PCR.* Biochim Biophys Acta, 1992. **1180**(2): p. 113-22.

99. Soong, N.W., et al., *Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain.* Nat Genet, 1992. **2**(4): p. 318-23.

100. Sugiyama, S., et al., *Quantitative analysis of age-associated accumulation of mitochondrial DNA with deletion in human hearts.* Biochem Biophys Res Commun, 1991. **180**(2): p. 894-9.

101. Wei, Y.H., *Mitochondrial DNA alterations as ageing-associated molecular events.* Mutat Res, 1992. **275**(3-6): p. 145-55.

102. Yen, T.C., et al., *Age-dependent increase of mitochondrial DNA deletions together with lipid peroxides and superoxide dismutase in human liver mitochondria.* Free Radic Biol Med, 1994. **16**(2): p. 207-14.

103. Yen, T.C., et al., *Age-dependent 6kb deletion in human liver mitochondrial DNA.* Biochem Int, 1992. **26**(3): p. 457-68.

104. Yen, T.C., et al., *Ageing-associated 5 kb deletion in human liver mitochondrial DNA.* Biochem Biophys Res Commun, 1991. **178**(1): p. 124-31.

105. Liu, V.W., C. Zhang, and P. Nagley, *Mutations in mitochondrial DNA accumulate differentially in three different human tissues during ageing.* Nucleic Acids Res, 1998. **26**(5): p. 1268-75.

106. Zhang, C., et al., *Multiple mitochondrial DNA deletions in an elderly human individual.* FEBS Lett, 1992. **297**(1-2): p. 34-8.

107. Zhang, C., et al., *Mitochondrial DNA deletions in human cardiac tissue show a gross mosaic distribution.* Biochem Biophys Res Commun, 1999. **254**(1): p. 152-7.

108. Zhang, C., et al., *Differential occurrence of mutations in mitochondrial DNA of human skeletal muscle during aging.* Hum Mutat, 1998. **11**(5): p. 360-71.

109. Zhang, J., et al., *The mitochondrial common deletion in Parkinson's disease and related movement disorders.* Parkinsonism Relat Disord, 2002. **8**(3): p. 165-70.

110. Liu, V.W., et al., *Independent occurrence of somatic mutations in mitochondrial DNA of human skin from subjects of various ages.* Hum Mutat, 1998. **11**(3): p. 191-6.

111. Zhang, C., A.W. Linnane, and P. Nagley, *Occurrence of a particular base substitution (3243 A to G) in mitochondrial DNA of tissues of ageing humans.* Biochem Biophys Res Commun, 1993. **195**(2): p. 1104-10.

112. Kadenbach, B., et al., *Human aging is associated with stochastic somatic mutations of mitochondrial DNA.* Mutat Res, 1995. **338**(1-6): p. 161-72.

113. Munscher, C., J. Muller-Hocker, and B. Kadenbach, *Human aging is associated with various point mutations in tRNA genes of mitochondrial DNA.* Biol Chem Hoppe Seyler, 1993. **374**(12): p. 1099-104.

114. Munscher, C., et al., *The point mutation of mitochondrial DNA characteristic for MERRF disease is found also in healthy people of different ages.* FEBS Lett, 1993. **317**(1-2): p. 27-30.

115. Linnane, A.W., et al., *Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases.* Lancet, 1989. **1**(8639): p. 642-5.

116. Wallace, D.C., et al., *Diseases resulting from mitochondrial DNA point mutations.* J Inherit Metab Dis, 1992. **15**(4): p. 472-9.

117. Wallace, D.C., *Mitochondrial DNA mutations in diseases of energy metabolism.* J Bioenerg Biomembr, 1994. **26**(3): p. 241-50.

118. Wallace, D.C., *Mitochondrial DNA in aging and disease.* Sci Am, 1997. **277**(2): p. 40-7.

119. Wallace, D.C., *A mitochondrial paradigm for degenerative diseases and ageing.* Novartis Found Symp, 2001. **235**: p. 247-63; discussion 263-6.

120. Linnane, A.W., et al., *Mitochondrial DNA mutation and the ageing process: bioenergy and pharmacological intervention.* Mutat Res, 1992. **275**(3-6): p. 195-208.

121. Manini, T.M., et al., *European ancestry and resting metabolic rate in older African Americans.* Eur J Clin Nutr. **65**(6): p. 663-7.

122. Tranah, G.J., et al., *Mitochondrial DNA variation in human metabolic rate and energy expenditure.* Mitochondrion, 2011.

123. Blanc, S., et al., *Energy requirements in the eighth decade of life.* Am J Clin Nutr, 2004. **79**(2): p. 303-10.

124. Blanc, S., et al., *Influence of delayed isotopic equilibration in urine on the accuracy of the (2)H(2)(18)O method in the elderly.* J Appl Physiol, 2002. **92**(3): p. 1036-44.

125. Prentice, A.M., et al., *Physical activity and obesity: problems in correcting expenditure for body size.* Int J Obes Relat Metab Disord, 1996. **20**(7): p. 688-91.

126. *Series WTR. Energy and Protein Requirements: Report of a Joint FAP/WHO/UNU Expert Consultation.* Geneva, Switzerland: World Health Organization, 1985.

127. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

128. Ng, P.C. and S. Henikoff, *Predicting the effects of amino acid substitutions on protein function.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 61-80.

129. Li, B., et al., *Automated inference of molecular mechanisms of disease from amino acid substitutions.* Bioinformatics, 2009. **25**(21): p. 2744-50.

130. Huang, L.S., et al., *Binding of the respiratory chain inhibitor antimycin to the mitochondrial bc1 complex: a new crystal structure reveals an altered intramolecular hydrogen-bonding pattern.* J Mol Biol, 2005. **351**(3): p. 573-97.

131. Muramoto, K., et al., *A histidine residue acting as a controlling site for dioxygen reduction and proton pumping by cytochrome c oxidase.* Proc Natl Acad Sci U S A, 2007. **104**(19): p. 7881-6.

132. Shinzawa-Itoh, K., et al., *Structures and physiological roles of 13 integral lipids of bovine heart cytochrome c oxidase.* Embo J, 2007. **26**(6): p. 1713-25.

133. http://genetics.bwh.harvard.edu/vt/dokuwiki/start.

134. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.

135. Price, A.L., et al., *Pooled association tests for rare variants in exon-resequencing studies.* Am J Hum Genet, 2010. **86**(6): p. 832-8.

136. Coon, K.D., et al., *Quantitation of heteroplasmy of mtDNA sequence variants identified in a population of AD patients and controls by array-based resequencing.* Mitochondrion, 2006. **6**(4): p. 194-210.

137. Quinlan, C.L., et al., *The mechanism of superoxide production by the antimycin-inhibited mitochondrial Q-cycle.* J Biol Chem. **286**(36): p. 31361-72.

138. Heher, K.L. and D.R. Johns, *A maculopathy associated with the 15257 mitochondrial DNA mutation.* Arch Ophthalmol, 1993. **111**(11): p. 1495-9.

139. Howell, N., et al., *Leber's hereditary optic neuropathy: the etiological role of a mutation in the mitochondrial cytochrome b gene.* Genetics, 1993. **133**(1): p. 133-6.

140. Huoponen, K., et al., *The spectrum of mitochondrial DNA mutations in families with Leber hereditary optic neuroretinopathy.* Hum Genet, 1993. **92**(4): p. 379-84.

141. Johns, D.R. and M.J. Neufeld, *Cytochrome c oxidase mutations in Leber hereditary optic neuropathy.* Biochem Biophys Res Commun, 1993. **196**(2): p. 810-5.

142. Johns, D.R. and M.J. Neufeld, *Cytochrome b mutations in Leber hereditary optic neuropathy.* Biochem Biophys Res Commun, 1991. **181**(3): p. 1358-64.

143. Savontaus, M.L., *mtDNA mutations in Leber's hereditary optic neuropathy.* Biochim Biophys Acta, 1995. **1271**(1): p. 261-3.

144. Fauser, S., et al., *Sequence analysis of the complete mitochondrial genome in patients with Leber's hereditary optic neuropathy lacking the three most common pathogenic DNA mutations.* Biochem Biophys Res Commun, 2002. **295**(2): p. 342-7.

145. Povalko, N., et al., *A new sequence variant in mitochondrial DNA associated with high penetrance of Russian Leber hereditary optic neuropathy.* Mitochondrion, 2005. **5**(3): p. 194-9.

146. Brown, M.D., et al., *Novel mtDNA mutations and oxidative phosphorylation dysfunction in Russian LHON families.* Hum Genet, 2001. **109**(1): p. 33-9.

147. Cai, W., et al., *Mitochondrial variants may influence the phenotypic manifestation of Leber's hereditary optic neuropathy-associated ND4 G11778A mutation.* J Genet Genomics, 2008. **35**(11): p. 649-55.

148. Campos, Y., et al., *Bilateral striatal necrosis and MELAS associated with a new T3308C mutation in the mitochondrial ND1 gene.* Biochem Biophys Res Commun, 1997. **238**(2): p. 323-5.

149. Efremov, R.G. and L.A. Sazanov, *Structure of the membrane domain of respiratory complex I.* Nature. **476**(7361): p. 414-20.

150. Roessler, M.M., et al., *Direct assignment of EPR spectra to structurally defined iron-sulfur clusters in complex I by double electron-electron resonance.* Proc Natl Acad Sci U S A. **107**(5): p. 1930-5.

151. Nijtmans, L.G., et al., *Impaired ATP synthase assembly associated with a mutation in the human ATP synthase subunit 6 gene.* J Biol Chem, 2001. **276**(9): p. 6755-62.

152. Jonckheere, A.I., et al., *A novel mitochondrial ATP8 gene mutation in a patient with apical hypertrophic cardiomyopathy and neuropathy.* BMJ Case Rep, 2009. **2009**.

153. MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org.
154. Terasaki, F., et al., *A case of cardiomyopathy showing progression from the hypertrophic to the dilated form: association of Mt8348A-->G mutation in the mitochondrial tRNA(Lys) gene with severe ultrastructural alterations of mitochondria in cardiomyocytes.* Jpn Circ J, 2001. **65**(7): p. 691-4.
155. Hirst, J., *Towards the molecular mechanism of respiratory complex I.* Biochem J. **425**(2): p. 327-39.
156. Langston, J.W. and P.A. Ballard, Jr., *Parkinson's disease in a chemist working with 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine.* N Engl J Med, 1983. **309**(5): p. 310.
157. Ramsay, R.R. and T.P. Singer, *Energy-dependent uptake of N-methyl-4-phenylpyridinium, the neurotoxic metabolite of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine, by mitochondria.* J Biol Chem, 1986. **261**(17): p. 7585-7.
158. Zimmermann, F.A., et al., *Respiratory chain complex I is a mitochondrial tumor suppressor of oncocytic tumors.* Front Biosci (Elite Ed). **3**: p. 315-25.
159. Haas, R.H., et al., *Low platelet mitochondrial complex I and complex II/III activity in early untreated Parkinson's disease.* Ann Neurol, 1995. **37**(6): p. 714-22.
160. Stefanatos, R. and A. Sanz, *Mitochondrial complex I: a central regulator of the aging process.* Cell Cycle. **10**(10): p. 1528-32.
161. Kathiresan, S., et al., *Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants.* Nat Genet, 2009. **41**(3): p. 334-41.
162. Kathiresan, S., et al., *Common variants at 30 loci contribute to polygenic dyslipidemia.* Nat Genet, 2009. **41**(1): p. 56-65.
163. Newton-Cheh, C., et al., *Genome-wide association study identifies eight loci associated with blood pressure.* Nat Genet, 2009. **41**(6): p. 666-76.
164. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
165. Ruiz-Pesini, E., et al., *Effects of purifying and adaptive selection on regional variation in human mtDNA.* Science, 2004. **303**(5655): p. 223-6.
166. Manini, T.M., *Energy expenditure and aging.* Ageing Res Rev, 2010. **9**(1): p. 1-11.
167. Manini, T.M., et al., *European ancestry and resting metabolic rate in older African Americans.* Eur J Clin Nutr, 2011. **65**(6): p. 663-7.
168. Donati, A., et al., *Age-related changes in the regulation of autophagic proteolysis in rat isolated hepatocytes.* J Gerontol A Biol Sci Med Sci, 2001. **56**(7): p. B288-93.
169. de Grey, A.D., *A proposed refinement of the mitochondrial free radical theory of aging.* Bioessays, 1997. **19**(2): p. 161-6.
170. Baur, J.A., et al., *Resveratrol improves health and survival of mice on a high-calorie diet.* Nature, 2006. **444**(7117): p. 337-42.
171. Davis, J.M., et al., *Quercetin increases brain and muscle mitochondrial biogenesis and exercise tolerance.* Am J Physiol Regul Integr Comp Physiol, 2009. **296**(4): p. R1071-7.

172. Liu, Z., et al., *Hydroxytyrosol protects retinal pigment epithelial cells from acrolein-induced oxidative stress and mitochondrial dysfunction.* J Neurochem, 2007. **103**(6): p. 2690-2700.

173. Rasbach, K.A. and R.G. Schnellmann, *Isoflavones promote mitochondrial biogenesis.* J Pharmacol Exp Ther, 2008. **325**(2): p. 536-43.

174. Stites, T., et al., *Pyrroloquinoline quinone modulates mitochondrial quantity and function in mice.* J Nutr, 2006. **136**(2): p. 390-6.

175. Chowanadisai, W., et al., *Pyrroloquinoline quinone stimulates mitochondrial biogenesis through cAMP response element-binding protein phosphorylation and increased PGC-1alpha expression.* J Biol Chem, 2010. **285**(1): p. 142-52.

176. Timmers, S., et al., *Calorie restriction-like effects of 30 days of resveratrol supplementation on energy metabolism and metabolic profile in obese humans.* Cell Metab. **14**(5): p. 612-22.

177. Guarente, L., *Mitochondria--a nexus for aging, calorie restriction, and sirtuins?* Cell, 2008. **132**(2): p. 171-6.

178. Civitarese, A.E., et al., *Calorie Restriction Increases Muscle Mitochondrial Biogenesis in Healthy Humans.* PLoS Med, 2007. **4**(3): p. e76.

179. Menshikova, E.V., et al., *Effects of exercise on mitochondrial content and function in aging human skeletal muscle.* J Gerontol A Biol Sci Med Sci, 2006. **61**(6): p. 534-40.

180. Johnston, A.P., M. De Lisio, and G. Parise, *Resistance training, sarcopenia, and the mitochondrial theory of aging.* Appl Physiol Nutr Metab, 2008. **33**(1): p. 191-9.

181. Hebert, L.E., et al., *Alzheimer disease in the US population: prevalence estimates using the 2000 census.* Arch Neurol, 2003. **60**(8): p. 1119-22.

182. Bishop, N.A., T. Lu, and B.A. Yankner, *Neural mechanisms of ageing and cognitive decline.* Nature, 2010. **464**(7288): p. 529-35.

183. Gibson, G.E. and Q. Shi, *A mitocentric view of Alzheimer's disease suggests multi-faceted treatments.* J Alzheimers Dis, 2010. **20 Suppl 2**: p. S591-607.

184. Nunomura, A., et al., *Oxidative damage is the earliest event in Alzheimer disease.* J Neuropathol Exp Neurol, 2001. **60**(8): p. 759-67.

185. Coskun, P.E., et al., *Systemic mitochondrial dysfunction and the etiology of Alzheimer's disease and down syndrome dementia.* J Alzheimers Dis. **20 Suppl 2**: p. S293-310.

186. Hirai, K., et al., *Mitochondrial abnormalities in Alzheimer's disease.* J Neurosci, 2001. **21**(9): p. 3017-23.

187. Pratico, D., et al., *Increase of brain oxidative stress in mild cognitive impairment: a possible predictor of Alzheimer disease.* Arch Neurol, 2002. **59**(6): p. 972-6.

188. Arlt, S., U. Beisiegel, and A. Kontush, *Lipid peroxidation in neurodegeneration: new insights into Alzheimer's disease.* Curr Opin Lipidol, 2002. **13**(3): p. 289-94.

189. Harris, M.E., et al., *Direct evidence of oxidative injury produced by the Alzheimer's beta-amyloid peptide (1-40) in cultured hippocampal neurons.* Exp Neurol, 1995. **131**(2): p. 193-202.

190. Lovell, M.A., et al., *Decreased thioredoxin and increased thioredoxin reductase levels in Alzheimer's disease brain.* Free Radic Biol Med, 2000. **28**(3): p. 418-27.

191. Pocernich, C.B. and D.A. Butterfield, *Acrolein inhibits NADH-linked mitochondrial enzyme activity: implications for Alzheimer's disease.* Neurotox Res, 2003. **5**(7): p. 515-20.

192. Yao, Y., et al., *Enhanced brain levels of 8,12-iso-iPF2alpha-VI differentiate AD from frontotemporal dementia.* Neurology, 2003. **61**(4): p. 475-8.

193. Chinopoulos, C., L. Tretter, and V. Adam-Vizi, *Depolarization of in situ mitochondria due to hydrogen peroxide-induced oxidative stress in nerve terminals: inhibition of alpha-ketoglutarate dehydrogenase.* J Neurochem, 1999. **73**(1): p. 220-8.

194. Park, L.C., et al., *Metabolic impairment induces oxidative stress, compromises inflammatory responses, and inactivates a key mitochondrial enzyme in microglia.* J Neurochem, 1999. **72**(5): p. 1948-58.

195. Gibson, G.E., et al., *Differential alterations in antioxidant capacity in cells from Alzheimer patients.* Biochim Biophys Acta, 2000. **1502**(3): p. 319-29.

196. Nulton-Persson, A.C., et al., *Reversible inactivation of alpha-ketoglutarate dehydrogenase in response to alterations in the mitochondrial glutathione status.* Biochemistry, 2003. **42**(14): p. 4235-42.

197. Humphries, K.M. and L.I. Szweda, *Selective inactivation of alpha-ketoglutarate dehydrogenase and pyruvate dehydrogenase: reaction of lipoic acid with 4-hydroxy-2-nonenal.* Biochemistry, 1998. **37**(45): p. 15835-41.

198. Rokutan, K., K. Kawai, and K. Asada, *Inactivation of 2-oxoglutarate dehydrogenase in rat liver mitochondria by its substrate and t-butyl hydroperoxide.* J Biochem, 1987. **101**(2): p. 415-22.

199. Correa, J.G. and A.O. Stoppani, *Catecholamines enhance dihydrolipoamide dehydrogenase inactivation by the copper Fenton system. Enzyme protection by copper chelators.* Free Radic Res, 1996. **24**(4): p. 311-22.

200. Hinerfeld, D., et al., *Endogenous mitochondrial oxidative stress: neurodegeneration, proteomic analysis, specific respiratory chain defects, and efficacious antioxidant therapy in superoxide dismutase 2 null mice.* J Neurochem, 2004. **88**(3): p. 657-67.

201. Pruijn, F.B., W.G. Schoonen, and H. Joenje, *Inactivation of mitochondrial metabolism by hyperoxia-induced oxidative stress.* Ann N Y Acad Sci, 1992. **663**: p. 453-5.

202. Schoonen, W.G., et al., *Characterization of oxygen-resistant Chinese hamster ovary cells. III. Relative resistance of succinate and alpha-ketoglutarate dehydrogenases to hyperoxic inactivation.* Free Radic Biol Med, 1991. **10**(2): p. 111-8.

203. Shi, Q., et al., *Novel functions of the alpha-ketoglutarate dehydrogenase complex may mediate diverse oxidant-induced changes in mitochondrial enzymes associated with Alzheimer's disease.* Biochim Biophys Acta, 2008. **1782**(4): p. 229-38.

204. Gabuzda, D., et al., *Inhibition of energy metabolism alters the processing of amyloid precursor protein and induces a potentially amyloidogenic derivative.* J Biol Chem, 1994. **269**(18): p. 13623-8.

205. O'Connor, T., et al., *Phosphorylation of the translation initiation factor eIF2alpha increases BACE1 levels and promotes amyloidogenesis.* Neuron, 2008. **60**(6): p. 988-1009.

206. Tamagno, E., et al., *Oxidative stress increases expression and activity of BACE in NT2 neurons.* Neurobiol Dis, 2002. **10**(3): p. 279-88.

207. Tamagno, E., et al., *Oxidative stress activates a positive feedback between the gamma- and beta-secretase cleavages of the beta-amyloid precursor protein.* J Neurochem, 2008. **104**(3): p. 683-95.

208. Tong, Y., et al., *Oxidative stress potentiates BACE1 gene expression and Abeta generation.* J Neural Transm, 2005. **112**(3): p. 455-69.

209. Calingasan, N.Y., K. Uchida, and G.E. Gibson, *Protein-bound acrolein: a novel marker of oxidative stress in Alzheimer's disease.* J Neurochem, 1999. **72**(2): p. 751-6.

210. Hauptmann, S., et al., *Mitochondrial dysfunction: an early event in Alzheimer pathology accumulates with age in AD transgenic mice.* Neurobiol Aging, 2009. **30**(10): p. 1574-86.

211. Yao, J., et al., *Mitochondrial bioenergetic deficit precedes Alzheimer's pathology in female mouse model of Alzheimer's disease.* Proc Natl Acad Sci U S A, 2009. **106**(34): p. 14670-5.

212. Hsiao, K., et al., *Correlative memory deficits, Abeta elevation, and amyloid plaques in transgenic mice.* Science, 1996. **274**(5284): p. 99-102.

213. Manczak, M., et al., *Mitochondria are a direct site of A beta accumulation in Alzheimer's disease neurons: implications for free radical generation and oxidative damage in disease progression.* Hum Mol Genet, 2006. **15**(9): p. 1437-49.

214. Mucke, L., et al., *High-level neuronal expression of abeta 1-42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation.* J Neurosci, 2000. **20**(11): p. 4050-8.

215. Pratico, D., et al., *Increased lipid peroxidation precedes amyloid plaque formation in an animal model of Alzheimer amyloidosis.* J Neurosci, 2001. **21**(12): p. 4183-7.

216. Leuner, K., et al., *Mitochondria-derived ROS lead to enhanced amyloid beta formation.* Antioxid Redox Signal.

217. de la Monte, S.M., et al., *Mitochondrial DNA damage as a mechanism of cell loss in Alzheimer's disease.* Lab Invest, 2000. **80**(8): p. 1323-35.

218. Anglade, P., et al., *Apoptosis in dopaminergic neurons of the human substantia nigra during normal aging.* Histol Histopathol, 1997. **12**(3): p. 603-10.

219. Dragicevic, N., et al., *Mitochondrial amyloid-beta levels are associated with the extent of mitochondrial dysfunction in different brain regions and the degree of cognitive impairment in Alzheimer's transgenic mice.* J Alzheimers Dis. **20 Suppl 2**: p. S535-50.

220. Gibson, G.E. and H.M. Huang, *Mitochondrial enzymes and endoplasmic reticulum calcium stores as targets of oxidative stress in neurodegenerative diseases.* J Bioenerg Biomembr, 2004. **36**(4): p. 335-40.

221. Haces, M.L., T. Montiel, and L. Massieu, *Selective vulnerability of brain regions to oxidative stress in a non-coma model of insulin-induced hypoglycemia.* Neuroscience, 2010. **165**(1): p. 28-38.
222. Eckert, A., et al., *Mitochondrial dysfunction, apoptotic cell death, and Alzheimer's disease.* Biochem Pharmacol, 2003. **66**(8): p. 1627-34.
223. Blass, J.P., G.E. Gibson, and S. Hoyer, *The role of the metabolic lesion in Alzheimer's disease.* J Alzheimers Dis, 2002. **4**(3): p. 225-32.
224. Grazina, M., et al., *Genetic basis of Alzheimer's dementia: role of mtDNA mutations.* Genes Brain Behav, 2006. **5 Suppl 2**: p. 92-107.
225. Castellani, R., et al., *Role of mitochondrial dysfunction in Alzheimer's disease.* J Neurosci Res, 2002. **70**(3): p. 357-60.
226. Corral-Debrinski, M., et al., *Marked changes in mitochondrial DNA deletion levels in Alzheimer brains.* Genomics, 1994. **23**(2): p. 471-6.
227. Lin, F.H., et al., *Detection of point mutations in codon 331 of mitochondrial NADH dehydrogenase subunit 2 in Alzheimer's brains.* Biochem Biophys Res Commun, 1992. **182**(1): p. 238-46.
228. Shoffner, J.M., et al., *Mitochondrial DNA variants observed in Alzheimer disease and Parkinson disease patients.* Genomics, 1993. **17**(1): p. 171-84.
229. Hutchin, T. and G. Cortopassi, *A mitochondrial DNA clone is associated with increased risk for Alzheimer disease.* Proc Natl Acad Sci U S A, 1995. **92**(15): p. 6892-5.
230. Egensperger, R., et al., *Association of the mitochondrial tRNA(A4336G) mutation with Alzheimer's and Parkinson's diseases.* Neuropathol Appl Neurobiol, 1997. **23**(4): p. 315-21.
231. Edland, S.D., et al., *Increased risk of dementia in mothers of Alzheimer's disease cases: evidence for maternal inheritance.* Neurology, 1996. **47**(1): p. 254-6.
232. Grazina, M., et al., *Mitochondrial DNA variants in a portuguese population of patients with Alzheimer's disease.* Eur Neurol, 2005. **53**(3): p. 121-4.
233. Qiu, X., Y. Chen, and M. Zhou, *Two point mutations in mitochondrial DNA of cytochrome c oxidase coexist with normal mtDNA in a patient with Alzheimer's disease.* Brain Res, 2001. **893**(1-2): p. 261-3.
234. Tanno, Y., K. Okuizumi, and S. Tsuji, *mtDNA polymorphisms in Japanese sporadic Alzheimer's disease.* Neurobiol Aging, 1998. **19**(1 Suppl): p. S47-51.
235. Brown, M.D., et al., *Mitochondrial DNA sequence analysis of four Alzheimer's and Parkinson's disease patients.* Am J Med Genet, 1996. **61**(3): p. 283-9.
236. Edland, S.D., et al., *Mitochondrial genetic variants and Alzheimer disease: a case-control study of the T4336C and G5460A variants.* Alzheimer Dis Assoc Disord, 2002. **16**(1): p. 1-7.
237. Kosel, S., et al., *No association of mutations at nucleotide 5460 of mitochondrial NADH dehydrogenase with Alzheimer's disease.* Biochem Biophys Res Commun, 1994. **203**(2): p. 745-9.
238. Petruzzella, V., X. Chen, and E.A. Schon, *Is a point mutation in the mitochondrial ND2 gene associated with Alzheimer's disease.* Biochem Biophys Res Commun, 1992. **186**(1): p. 491-7.

239. Wragg, M.A., et al., *No association found between Alzheimer's disease and a mitochondrial tRNA glutamine gene variant.* Neurosci Lett, 1995. **201**(2): p. 107-10.

240. Tysoe, C., et al., *The tRNA(Gln) 4336 mitochondrial DNA variant is not a high penetrance mutation which predisposes to dementia before the age of 75 years.* J Med Genet, 1996. **33**(12): p. 1002-6.

241. Janetzky, B., et al., *Investigations on the point mutations at nt 5460 of the mtDNA in different neurodegenerative and neuromuscular diseases.* Eur Neurol, 1996. **36**(3): p. 149-53.

242. Teng, E.L. and H.C. Chui, *The Modified Mini-Mental State (3MS) examination.* J Clin Psychiatry, 1987. **48**(8): p. 314-8.

243. Wechsler, D., *Wechsler Adult Intelligence Scale - Revised. San Antonio, Psychological Corporation.* 1981.

244. Beres, C.A. and A. Baron, *Improved digit symbol substitution by older women as a result of extended practice.* J Gerontol, 1981. **36**(5): p. 591-7.

245. Tierney, M.C., et al., *Psychometric differentiation of dementia. Replication and extension of the findings of Storandt and coworkers.* Arch Neurol, 1987. **44**(7): p. 720-2.

246. Fiocco, A.J., et al., *COMT genotype and cognitive function: an 8-year longitudinal study in white and black elders.* Neurology. **74**(16): p. 1296-302.

247. Patel, J., R.R. Pathak, and S. Mujtaba, *The biology of lysine acetylation integrates transcriptional programming and metabolism.* Nutr Metab (Lond). **8**: p. 12.

248. Zhao, S., et al., *Regulation of cellular metabolism by protein lysine acetylation.* Science. **327**(5968): p. 1000-4.

249. Wang, Q., et al., *Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux.* Science. **327**(5968): p. 1004-7.

250. Yu, W., et al., *Lysine 88 acetylation negatively regulates ornithine carbamoyltransferase activity in response to nutrient signals.* J Biol Chem, 2009. **284**(20): p. 13669-75.

251. Lee, J.W., et al., *Mitochondrial DNA copy number in peripheral blood is associated with cognitive function in apparently healthy elderly women.* Clin Chim Acta. **411**(7-8): p. 592-6.

252. Migliore, L., et al., *Oxidative DNA damage in peripheral leukocytes of mild cognitive impairment and AD patients.* Neurobiol Aging, 2005. **26**(5): p. 567-73.

# Chapter 3

# Discovery of variants underlying chemotherapy response by pathway-based candidate gene resequencing

## *3.1 Abstract*

The realization of personalized medicine relies on a comprehensive understanding of how genetic variants contribute to drug response. Much like other complex traits, drug response is multigenic in nature; therefore, effectively identifying relevant variants requires high-throughput, genome-scale approaches. The advent of massively parallel sequencing or next-generation sequencing technologies promises to enable a wide range of applications such as resequencing, *de novo* sequencing, expression profiling, and regulatory sequence surveys in an ultra high-throughput fashion and at a much lower cost (reviewed in [1]). We hypothesize that functionally relevant variants in drug pathway genes contribute to extreme drug response phenotypes. Specifically, I will focus on discovery of variants associated with chemotherapy response for developing a framework potentially applicable to other drug classes. I will describe the use of emulsion PCR and next-generation sequencing to sequence 95 drug pathway candidate genes in DNA from 95 cell lines with known drug response phenotypes.

## *3.2 Introduction*

Drug response is a complex trait, polygenic and multifactorial. Environmental factors such as drug-drug interactions and the nature of disease undoubtedly contribute to the overall drug response phenotype. However, drug response is heritable and has a genetic component [2]. Therefore, the drug response phenotype will reflect a complex interplay of both genetic and non-genetic factors.

### *3.2.1 Genetic basis of drug response phenotypes*

Genomics has the potential to be a valuable aid in bringing personalized medicine closer to reality [3]. The U.S Food and Drug Administration also recognizes the importance of genetic biomarkers and now maintains a database documenting well-supported polymorphisms that contribute to drug toxicity or resistance ("Table of Valid Genomic Biomarkers in the Context of Approved Drug Labels"; http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm). The two goals in pharmacogenetics are to identify responders and non-responders, and to identify those at risk of having serious adverse reactions (ADRs). However, genotype-informed individualized therapy requires knowledge of the genes a given drug interacts with, genetic variability of those genes, and the effects of the variation on efficacy and toxicity. The ability to predict chemotherapy response from genotype is particularly important, since most of these agents have narrow therapeutic windows.

While potentially curative, surgical resection of tumor is not suitable for all cancer patients. Therefore, chemotherapy agents are used for treatment of cancer in conjunction with or in place of surgery. There is, however, inter-individual variation in drug response. Both toxicity and lack of efficacy are undesirable but common in the clinic. Traditional methods of adjusting dose based on body surface area measurements and weight have limited predictive power and are clearly inadequate [4]. Inter-individual variation in response to cytotoxic agents traditionally used in chemotherapy could be attributed partly to a genetic component. For example, the estimated heritability for 5-fluorouracil cytotoxicity ranges from 0.26 to 0.65 [5]. A number of variants in a myriad of genes have been identified to be associated with drug response. Many variants affecting drug response are expected to have little phenotypic effect except during drug administration. Because these chemical entities were introduced very recently in the human evolutionary history, variants that bear major effects on drug response may be maintained in the population. Therefore, the identification of these variants that are relevant to drug response is crucial.

Toxicity and resistance represent two extreme ends of the drug response spectrum. Many drug-metabolizing enzymes (such as Phase I and II enzymes) and transporters (such as ABC and SLC transporters) could interact with multiple drugs, so one would expect variation in these genes to broadly affect drug response, potentially causing either toxicity or resistance. At the same time, each drug interacts with a unique set of pharmacodynamic drug targets.

### *3.2.2 Chemotherapy agents of interest*

A new generation of chemotherapy agents that inhibit specific targets uniquely expressed in cancer cells is being developed. A noted example is imatinib (marketed by Novartis under the brand name Gleevec), which targets a fusion oncogene product, BCR-ABL, found in chronic myeloid leukemia cells. However, most common chemotherapy agents are general cytotoxic agents. They often interfere with replication and the cell cycle. The premise of using general cytotoxic agents is that although these drugs are expected to have a system-wide effect, cancer cells are the most adversely affected because of the observation that they are more rapidly dividing. Cytotoxic agents have been widely used and are effective in at least a subset of patients. Due to their system-wide effects, these drugs typically have narrow therapeutic windows. Toxicities and resistance have been described in the five drug classes being considered in this study.

### *5-Fluorouracil (5-FU)*

5-FU, which belongs to the drug class antimetabolites, has been in use for a range of cancers, including colorectal, pancreatic, and breast cancers. It inhibits biosynthetic processes by interfering with the nucleotide synthetic enzyme thymidylate synthase and by being mis-incorporated into RNA. Numerous studies involving the use of 5-FU as a single-agent treatment or in combination with other chemotherapy agents have been conducted to assess its efficacy; however, the response rates reported are highly variable (reviewed in [6]).

Encoded by *TYMS* , thymidylate synthase (TS) provides the sole *de novo* source of thymidine, so it serves a pivotal role in the biosynthesis of DNA. Competitive binding of 5-FU to thymidylate synthase blocks synthesis of the pyrimidine thymidine required for DNA synthesis and repair. Thymidylate synthase methylates deoxyuridine monophosphate (dUMP) into thymidine monophosphate (dTMP), which is then phosphorylated to thymidine triphosphate. Increased *TYMS* gene expression levels [7] and variants in *TYMS* [8] have been shown to be predictive of poor response or resistance to chemotherapy with 5-FU.

Dihydropyrimidine dehydrogenase (DPD) catabolizes 5-FU into dihydrofluorouracil (DHFU). An early case study documents a patient with undetectable DPD activity who had a severe adverse drug reaction [9]. More recently, common genetic variants in *DPYD* (which encodes for DPD) have been shown to be associated with 5-FU toxicities in larger patient cohorts [10].

Myriad Genetic Laboratories commercializes TheraGuide 5-FU$^{TM}$, marketed as a "test that allows you to customize management and minimize your patients' risk for an adverse reaction to 5-FU-related chemotherapy". The test genotypes common variants in *DPYD* and *TYMS* that have been implicated in 5-FU-related toxicities (http://www.myriadtests.com/hcp/about_theraguide.php).

***Camptothecins***

A quinoline alkaloid initially isolated from the tree *Camptotheca acuminata*, camptothecin (CPT) acts as an inhibitor of DNA topoisomerase I (TOP1) [11]. Because of TOP1's role in relaxing DNA supercoiling prior to replication and transcription, its inhibition leads to DNA breakage and eventually apoptosis. Two FDA-approved CPT analogues are in use – topotecan for ovarian, lung, and cervical cancer, and irinotecan for colorectal cancer.

Irinotecan is a prodrug and relies on hydrolysis to be converted to the active metabolite SN-38. SN-38 is inactivated through glucuronidation by UGT1A1, in which polymorphisms such as the *28 allele have been found to be associated with toxicity [12]. The FDA has in fact changed irinotecan's labeling and recommended genotype consideration.

A number of nonsynonymous variants were found in TOP1 [13]. Interestingly, the N722S variant was also found in CPT producing plants, prompting the authors to suggest that the variant offers protection to these plants [14].

### *Platinum*

A number of platinum-based agents have shown efficacy against a range of tumors such as colorectal, bladder, and cervical cancer [15]. Notable examples in this drug class include cisplatin and carboplatin. They have the general form *cis*-$[PtX_2(NHR_2)_2]$, where X is a leaving group and R is an organic fragment. These compounds are first hydrated, giving rise to the active species $Pt^{2+}(NHR_2)_2$, which then binds to two guanine bases. Intrastrand and interstrand cross-linking creates local distortion of the helical structure and interferes with DNA replication and transcription [16].

Much research has focused on the role of DNA repair pathways on platinum response [17], since DNA repair mechanisms are responsible for recognition and removal of these platinum-DNA adducts. In a study, approximately 90% of the patients with testicular cancer achieved cure by cisplatin-based therapy [18]. It was later shown that testicular tumors express low levels of XPA and ERCC1-XPF, which are genes in the nucleotide excision pathway [19, 20]. Other mechanisms of resistance have also been reviewed in [21].

### *Taxanes*

The first taxanes introduced were natural products. Taxane analogues are used to inhibit cell cycle progression. They hyperstabilize microtubules by binding to the β-subunit of tubulin, preventing restructuring of microtubules needed for proper cell division. Both docetaxel and paclitaxel belong to the family of taxanes. Toxicity (especially hematological toxicities) is common but manageable in patients on taxane monotherapy [22]. Resistance has been attributed partly to variants found in β-tubulin [23] and microtubule-associated proteins [24]. However, in a large study with 914 ovarian cancer patients, no association was found between outcome or toxicity and variants in 11 pathway genes including selected ABC transporters, P450, and microtubule-associated proteins [25]. Several smaller studies did demonstrate an effect of variants of ABCB1 on taxane response [26, 27].

### *3.2.3 Resequencing for rare variant discovery*

Most examples of genotype-phenotype association in the literature involve common polymorphisms whose minor allele frequencies (MAFs) exceed 5% in a population. The usefulness of genome-wide association studies (GWAS) in identifying common single-nucleotide polymorphisms (SNPs) that play a role in disease susceptibility has been clearly demonstrated. However, for most complex traits, more than 90% of heritability remains unexplained by these common variants [28]. The observation has prompted a shift from focusing on common variants to rare variants with MAFs less than 5%. The rare variant hypothesis states that different low frequency variants in different genes act independently to confer a phenotype (reviewed in [29] and [30]). Most related studies have focused on identifying rare variants that contribute to disease susceptibility. As an example, multiple functional variants increase the risk of colorectal adenoma [31]. In the group of patients where the rare variants were identified, none had more than one variant, suggesting that a single variant could confer significant risk and that variants act independently.

Current SNP tagging methods aim to minimize the number of SNPs needed to be genotyped by maximizing the number of variants each genotyped SNP tags. However, because of their low minor allele frequencies, rare variants are by default in low LD with common SNPs, and hence, not covered well in genotyping arrays. The lack of coverage of rare variants necessitates resequencing.

### *3.2.4 Enrichment of target regions*

Whole genome sequencing using the next-generation sequencing platforms is now relatively common. Smaller genomes such C. elegans and Arabidopsis were sequenced initially, but several human genomes have been sequenced more recently as well [32-36]. The high throughput ability of these platforms is well-suited for whole-genome sequencing, but is not immediately applicable to studies where specific regions of the genome are of interest.

PCR is a well-established technique for amplifying specific regions. Traditional PCR allows amplification of target regions with high specificity and sensitivity. PCR primers can be designed against a divergent region among homologous sequences. Therefore, even coverage is possible across repetitive regions that hybridization-based methods typically miss. Also, primer design rules have been formulated and experimentally validated, enabling specific target enrichment.

However, PCR is labor-intensive and not amenable to high degrees of multiplexing. RainDance Technologies (RDT) commercialized a sequence enrichment platform base on an emulsion PCR approach. This RainStorm$^{TM}$ droplet-based approach could overcome the difficulties of multiplexing. By mixing the DNA template with different primer pairs in separate oil droplets, millions of single-plex PCR reactions can be performed simultaneously in a single tube.

The RainDance system's throughput is best suited for following up GWAS signals (including coding and non-coding regions one can design primers for) and studying rare variants in a set of regions. Their system produces 10 million homogeneous picoliter-size droplets per hour, each containing a single PCR reaction. Thus, while the approach allows for massively parallel amplification of a large number of targets, the specificity of any PCR reaction is not negatively affected because each amplicon is physically isolated [37]. RDT has developed a primer design pipeline to efficiently design primers for droplet libraries to target megabases of sequence. For example, RDT offers an oncology panel that targets 142 genes associated with cancer. The design includes 4,000 primer pairs that target coding exons, 5' and 3' splice sites, 3'UTR, and promoter regions. The amplicons often range from 300-500 bp. More recently, RDT has supported longer amplicons of up to a few kilobases as well, allowing for more flexibility in the primer design.

### *3.2.5 Central aim*

By using emulsion PCR to selectively amplify 95 drug pathway genes and next-generation sequencing to sequence the resulting amplicons, we aim to identify functionally relevant variants that contribute to extreme drug response phenotypes.

### *3.3 Methods*

#### ***Drug exposure and cell cytotoxicity measurement***

Dose-response curves were generated for each drug for a subset of the cell lines and expected to be sigmoidal. A specific dose that is closest to the $EC_{50}$ of the drugs was chosen to be applied to all cell lines. Cell viability had to be around 90% as determined by a standard trypan blue assay prior to the seeding of cells. Alamar blue, which assesses vital mitochondrial function, was used for measuring cytotoxicity. After Alamar blue was added, the cells were exposed to an excitation wavelength of 570 nm, and the emission at 600 nm was measured. $EC_{50}$ values were determined by measuring cell viability after different concentrations of the drugs were introduced.

#### ***Primer library generation.***

Primer pairs were designed based on Primer3 and synthesized to amplify the target genes. The primers were provided as a mixture at a concentration of 1.1 uM in 8 pL droplets by RDT. The droplets were checked for size and uniformity, and each primer pair was represented by an equal number of droplets.

#### ***Emulsion PCR***

2.5 ug of DNA were fragmented by hydroshearing (Digilab HydroShear) to 2-4 kb. The fragmented DNA was mixed with 2.5 uL of High-Fidelity buffer (10X; Invitrogen), 1.6 uL of $MgSO_4$, 0.8 uL 10 mM dNTP, 4.0 uL betaine, 4.0 uL of RDT Droplet Stabilizer, 2.0 uL dimethyl sulfoxide, and 0.8 uL of Platinum High-Fidelity Taq (5 U/uL). 20 uL of the DNA droplet mix were put on RDT1000 for droplet merging. The merged droplets underwent PCR amplification with the following PCR program: $94^0$C for 2 minutes; 55 cycles at $94^0$C for 15 seconds, $68^0$C for 30 seconds; $68^0$C for 10 minutes. The emulsion was broken by RDT Droplet Destabilizer. The mixture was vortexed for 15 seconds and spun in a microcentrifuge for 10 minutes.

### ***Sequencing sample preparation***

*Fragment end-repair*. 38 uL of DNA were mixed with 10 uL of End Repair Buffer (5X) and 2 uL of the End Repair Enzyme (EndTerminator kit; Lucigen). The mixture was incubated for 30 minutes at room temperature. The DNA was then bead-purified (as described in the last section) and eluted in 28 uL of EB.

*Concatenation of PCR products.* 28 uL of the end-repair reaction were mixed with 2 uL of T4 PNK, 2 uL of Quick Ligase, 50 uL of Quick Ligase Buffer (2X), and 18 uL of PEG 8000 (40%). The reaction was incubated for 15 minutes at room temperature. The DNA was then bead-purified and eluted in 38 uL of EB.

*Sonication and end-repair.* 10 uL of End Repair Buffer (5X) was added to 38 uL of the concatenated PCR products. The mixture was transferred to a Covaris tube. The DNA was fragmented to 100-600 bp (with an average of ~350 bp) by Covaris. 2 uL of End Repair enzyme was added directly to the Covaris tube with the DNA. The reaction was incubated for 30 minutes at room temperature, and then bead-purified and eluted in 32 uL of EB. For quality control, 2 uL of DNA from before and after the fragmentation were loaded onto a 2% agarose gel to verify the size distribution.

*A-addition*. 32 uL of the DNA from the end repair reaction were mixed with 5 uL of Klenow Buffer (10X), 10 uL of 1 mM dATP, and 3 uL of Klenow exo- (5U/uL). The reaction was incubated for 30 minutes in a thermocycler at $37^0$C, bead-purified, and eluted in 10 uL of EB.

*Adapter ligation*. 10 uL of the DNA from the A-tail reaction were mixed 25 uL of DNA ligase buffer, 5 uL of the Illumina adapter oligo mix, 5 uL of DNA ligase, and 5 uL of water. The reaction was incubated for 15 minutes at room temperature, bead-purified, and eluted in 20 uL of EB.

*PCR Amplification*. To optimize PCR conditions, a test PCR was set up using 1 uL of adapter-ligated DNA, 25 uL of Phusion Polymerase Master Mix, 23 uL of water, 0.5 uL of Illumina PE Primer 1.0 (25 uM), and 0.5 uL of Illumina PE Primer 2.0 (25 uM). 4 uL of the PCR mix was removed from the PCR tube at 2, 4, 6, 8, 10 cycles and visualized on a 2% agarose gel. 4 more PCR reactions were set up for each sample as above with the optimal number of cycles. The PCR products were bead-purified and eluted in 20 uL of EB.

*Bead purification*. The equivalent of 1.5X reaction volume of AMPure beads was added to the DNA sample. The mixture was incubated for 5 minutes at room temperature on a rotator. The beads were pelleted against the wall of the tube using a magnetic stand (Invitrogen). The supernatant was removed and the beads washed twice with 500 uL of 70% ethanol with a 30-second incubation. All supernatant was removed and the beads were dried in a speed-vac for 5 minutes at $37^0$C. The beads were resuspended in EB Buffer (pH 8.0; Qiagen). The beads were pelleted using the magnetic stand, and the supernatant containing purified DNA was transferred to a fresh tube.

### *Data generation and analysis*

Sequencing data of 35-40 bases were generated on the Illumina $GAII_x$ using the Illumina pipeline v1.3. Sequencing reads were aligned to the human genome reference (hg19) using Bowtie [38]. Only unique alignments were kept for analysis; they were allowed to have up to 2 mismatches. Post-processing was performed using samtools [39] (alignment merging), Picard (alignment sorting), and GATK [40] (realignment and quality score recalibration). Multi-sample variant calling was performed using GATK's Unified Genotyper with default settings.

## *3.4 Results*

### *Identification of extreme responders*

We have obtained cytotoxicity data by exposing lymphoblastoid cell lines (LCLs) derived at Centre d'Etude du Polymorphisme Humain (CEPH) to each chemotherapy agent under a standard set of conditions and subsequently measuring cell survival. As expected, response to these drugs was highly variable (Figure 3.1). We have chosen all the cell lines whose response to at least one of the drugs was on either tail (<10th percentile or >90th percentile) of the response distribution for our study in order to maximize statistical power [41]. We will focus on a total of 95 cell lines; a subset of them is either resistant or sensitive to multiple agents.

### *Emulsion PCR amplicon design*

In collaboration with RainDance Technologies, we have designed primers based on human reference hg18 to amplify promoter and exonic regions of candidate genes from five drug pathways (the genes, their functions and associated pathway(s) are summarized in Table 3.1 and 3.2). We compiled the gene lists based on PharmGKB drug pathways (http://www.pharmgkb.org/), which take into account of both pharmacokinetic interactions (how the drug enters the cell and reaches its target, and how it is eliminated) and pharmacodynamic interactions (how the drug exerts its cellular effects). Most genes are unique to the particular pathway they are part of. However, there is some overlap between pathways; notably, many are transporters assumed to be important for transport of multiple drugs (Figure 3.2).

The primers were designed away from known SNPs to minimized allelic imbalanced amplification. Our final design included 95 genes covered by 1932 amplicons. The primers mainly target exons and promoter regions of the target genes.

The amplicons ranged from 206-600 bp (mean = 514 bp) and had %GC from 24-78% (mean = 46%). 2.9% of the amplicons were predicted to have potential off-target effects.

***Targeted next generation sequencing***

Microdroplet emulsion PCR involves electrically merging droplets containing fragmented DNA with droplets containing primer pairs (Figure 3.3), multiplex thermal cycling of the PCR reactions, and breaking the emulsion to release of the PCR products (Figure 3.4). Prior to droplet merging, the genomic DNA was fragmented via hydroshearing to 2-4 kb in order to minimize off-target effects. Droplet merging was performed on the RDT 1000 at Washington University in St. Louis using RDT disposable microfluidic chips. The DNA does not interact with the walls of the microfluidic channels, so contamination is prevented. Droplet merging for each sample took approximately 40 minutes; the samples then underwent 34 cycles of PCR. This was followed by concatenation and fragmentation of the PCR amplicons so that the sequencing reads would have unique ends.

Sample preparation for Illumina sequencing was done at UCSF. Each sample was run on a single lane of the Illumina GAII$_x$; data were collected at the UCSF Genomics Core Facility (single-end 36-base), UCD Genome Center (single-end 40-base), and Illumina Services (single-end 35-base). The contributions by the three data sources are summarized in Table 3.3. As shown, the majority of the data was collected at UCSF.

The samples received an average of 22 million reads, and each sample had at least eight million reads. We developed a pipeline that automates sequence alignment and variant calling. To minimize misalignment, we only considered unique alignments (the alignment results are summarized in Figure 3.5). On average, 71% of the reads could be mapped uniquely to the human references sequence (hg19). Each sample received at least five million uniquely mapped reads (the mean was 15 million reads).

On average, 93.8% of the target bases received at least 15X coverage across the samples. This translates to a mean coverage of 288X across the target bases, which was deemed sufficient for high confidence variant calling. Sample 7029 had the worst enrichment performance, even though most of the reads from this sample could be uniquely mapped to the human genome, which indicated that the DNA was of adequate quality. It received an average coverage of 7.1X across target bases, and only 1.1% of the bases were covered at >15X coverage. Sample 7029 was excluded from further analysis. After the exclusion, the sample with the lowest coverage across target regions had an average coverage of 97X and 87.1% of the bases were covered at >15X coverage.

### *Variant analysis*

We performed multi-sample variant calling by GATK using the entire dataset. In total, we obtained 162112 high-confidence variants (defined as a position for which a non-reference allele is observed). First, we annotated the variants based on whether they were in genic and/or exonic regions using ANNOVAR. 82874 (51.12%) of them were intergenic, likely coming from sequencing of the background genomic DNA. 2481 were exonic, 532 were from 5'-UTR, and 1891 were from 3'-UTR. We also checked our variant list to determine how many of them had already been previously reported. Of all the variants, 102766 (63.4%) were already in dbSNP (v130) and had an associated rs number.

We employed various strategies to prioritize the variants found. Functional prediction involves assessing the degree of conservation and making inferences of whether a variant would have an impact on protein function. We used both PolyPhen (http://genetics.bwh.harvard.edu/pph/) and SIFT predictions. PolyPhen makes its predictions based on protein structure considerations, the degree of conservation at that position among homologous proteins, and whether the substitution falls within features in a protein. Similarly, SIFT predicts how well a substitution would be tolerated by constructing a multiple sequence alignment of related proteins and calculating the probabilities of possible substitutions [42]. As discussed in the next paragraphs, several variants were deemed of interest for future follow-up. For ease of reading, discussion of the variants is included in these results paragraphs.

### ***Response to topoisomerase inhibitors***

To identify variants associated with response to topoisomerase inhibitors, we combined the data for teniposide and topotecan. For initial filtering, we considered two individuals resistant to both drugs and compared them with four individuals sensitive to both drugs. Both resistant individuals and none of the sensitive individuals were homozygous for the wild-type allele (T) for the nonsynonymous variant rs1130609 (in *RRM2*), which leads to an amino acid change from serine to alanine. To examine the frequency of the alleles for this variant in our entire dataset, we considered individuals who were either resistant or sensitive to at least one of the two drugs. 11 out of 16 (69%) sensitive individuals versus 5 out of 19 (23%) resistant individuals were homozygous for the variant allele (G). SIFT predicted this non-synonymous change to be deleterious;

however, the variant allele is quite common in the Caucasian population (frequency of 0.63 in the 1000 Genomes samples). *RRM2* encodes for one of the subunits for ribonucleotide reductase, which catalyzes the generation of deoxyribonucleotides [43]. Since topoisomerase inhibitors promote double strand breaks and disrupt DNA replication, this variant could potentially contribute to differential drug response by modulating the rate of DNA synthesis and DNA repair.

### *Response to anti-microtubule agents*

Similar to what has been described in the previous section, we considered data from docetaxel, etoposide, paclitaxel, vincristine, and vinorelbine together. For discovery, we compared an individual resistant to all five agents with five individuals sensitive to all five agents. All sensitive individuals were homozygous for the wildtype allele for C/T splicing variant rs776746 (in *CYP3A5*); the resistant individual was heterozygous at the position. When considering the entire dataset, we found that most of the individuals with an extreme response to any of the five agents were homozygous for the wildtype allele. However, all four individuals that were heterozygous at that position were classified as being resistant to these agents. None of the samples were homozygous for the variant allele. This variant in the drug-metabolizing enzyme CYP3A5 has been well-studied. The wildtype allele is the minor allele; it has a frequency of 0.37 in the population (estimated based on the 1000 Genomes samples). Interesting, it creates a non-functional protein [44]. In our case, the variant represents a plausible candidate because of the hypothesis that as the variant encodes for a functional protein, metabolism of the drug is enhanced, leading to drug resistance. This is consistent with the fact that individuals with the variant allele were all classified as being resistant to anti-microtubule agents, all of which have been shown to be metabolized to some degree by CYP3A5.

### ***Response to antimetabolites***

To not rely on a discovery set of a few samples, we took a different approach when analyzing antimetabolite response. For this analysis, we used only exonic, nonsynonymous variants predicted by SIFT to be functional. We combined data from 5'-fluorouracil, cytarabine, fludarabine, and gemcitabine and focused on individuals with an extreme drug response to any of the four drugs. For a G/A variant in *MTHFR* rs1801133, six out of the 23 (35%) sensitive individuals were homozygous for the wildtype allele. In contrast, 15 out of the 30 (50%) resistant individuals were homozygous for the wildtype allele. The variant changes the amino acid from alanine to valine. Despite the seemingly innocuous change, SIFT predicted the change to functionally impact the protein. In fact, this position is highly conserved (LOD = 122). The estimated variant allele frequency ranges from 0.27 (based on exome sequencing data) to 0.4 (based on 1000 Genomes data). MTHFR is important for *de novo* thymidine and DNA and RNA synthesis [45]. Some of the individuals likely were more resistant to antimetabolite drugs because they had functional MTHFR.

### ***Multi-drug resistant/sensitive phenotypes***

It was noted that several individuals were either resistant or sensitive to multiple drugs. Therefore, we were interested in looking for variants that might contribute to this multi-drug resistant/sensitive phenotype. We classified each individual as being broadly resistant or sensitive to multiple agents (only individuals showing an extreme phenotype for at least 50% of the 20 drugs were kept for analysis) and looked for differential allele distribution among the resistant versus the sensitive individuals. For rs2228527 (in *ERCC6*), 20 out of the 25 (80%) multi-drug sensitive individuals had the wildtype allele; only 22 out of the 39 (56%) multi-drug resistant individual did. The variant is predicted to impact the protein's function by both SIFT and PolyPhen, despite the nucleotide position being not particularly conserved. Nonetheless, ERCC6 is involved in the DNA excision repair process [46], so it is natural to posit that variants affecting the function of ERCC6 may in turn affect a cell's ability to repair DNA damage when under the exposure of chemotherapy agents.

### *3.5 Discussion*

Paired with next generation sequencing, emulsion PCR has proven to be a high throughput way of targeting a large set of gene regions for our candidate gene study. This resequencing study has allowed us to study both common and rare variants in "high-value" regions of the genome.

Instead of using patient response data, our phenotyping was based on cell line data. *Ex vivo* models such as lymphoblastoid cell lines (LCLs) derived at Centre d'Etude du Polymorphisme Humain (CEPH) have been a valuable resource for pharmacogenetic studies. Several studies using LCLs from CEPH pedigrees have shown that LCL drug response phenotypes are heritable [5, 47, 48], validating the use of LCLs for pharmacogenetic studies. The use of LCLs eliminates environmental variability prevalent in clinical samples. It represents an unlimited source of DNA from normal subjects. It bypasses the obvious ethical issues with testing chemotherapy agents on normal subjects or even unaffected family members of patients with resistant or sensitive phenotypes.

However, the reliance on cell line data is not without potential flaws. First, specific chromosomal aberrations in LCLs have been documented [49]. However, these artifacts were observed in only a small number of samples and in a small percentage of the genome, and so they are assumed to have minimal impact on our study. Second, LCLs came from a single tissue type and regulatory variants impact gene expression in a cell type-dependent manner [50], so tissue-specific responses cannot be examined. Third, most CEPH LCLs were derived from Caucasian individuals; therefore, the findings may not apply to other populations. In fact, there is an abundance of data showing differential drug response in different populations. Wilson et al. showed previously that populations differ in their allelic frequencies of drug metabolizing enzymes [51]. Even so, we believe that establishing a catalogue of functionally important variants will be useful. Also, there are examples where a variant is associated with response across multiple populations [52].

One should consider some of the complications that using clinical materials may entail as well. In actual patients, there is considerable variation in severity of symptoms, compliance, and other environmental factors that confound the phenotype. Data or samples are often collected on the basis of convenience. Because of difficulties in identifying cases, most studies are limited to working with modest sample sizes.

There are several approaches one can take to follow up interesting findings. For demonstrative purposes, I will highlight *in vitro* studies that can be completed in a shorter time frame. For genes involved in the pharmacokinetics of a drug such as drug transporters (ABC and SLC) and drug metabolizing enzymes (CYP P450), variants can be expressed in cells and kinetic studies can be performed. One may expect a change in the pharmacokinetic profile if the enzyme metabolizes the drug of interest at a lower rate. One could also look at variants in pharmacodynamics targets. For example, paclitaxel works by binding to the microtubules, which are composed of TUBA and TUBB. There are also ancillary proteins such as MAP2, MAP4, and MAPT that assist in the reconstruction of microtubule needed for cell division. Variants can be cloned into a GFP-containing plasmid. Mutations in MAP2, MAP4, MAPT may affect their proteins' interactions with microtubules. Therefore, immunofluorescence experiments will be useful for studying the localization of these proteins. Variants in the proteins involved in microtubule reconstruction may affect the structure and formation of microtubules.

In summary, we identified variants that potentially contribute to drug resistance or toxicity in patients. The results may have clinical significance, and our findings can be useful in guiding genotype-informed therapy. This is the one of the first demonstration of the utility of emulsion PCR-based high-throughput targeted sequencing of drug pathway genes in a pharmacogenetic study, and this framework could be applied to other drugs with known pathways.

### *3.6 Funding sources*

The project was part of the **C**omprehensive **R**esearch on **E**xpressed **A**lleles in **T**herapeutic **E**valuation (CREATE) project and was funded by as one of the Pharmacogenetics Research Network grants U01 GM63340.

### *3.7 References*

1. Tucker, T., M. Marra, and J.M. Friedman, *Massively parallel sequencing: the next big thing in genetic medicine.* Am J Hum Genet, 2009. **85**(2): p. 142-54.
2. Evans, W.E. and H.L. McLeod, *Pharmacogenomics -- Drug Disposition, Drug Targets, and Side Effects.* N Engl J Med, 2003. **348**(6): p. 538-549.
3. Lord, P.G. and T. Papoian, *Genomics and Drug Toxicity.* Science, 2004. **306**(5696): p. 575-.
4. Gao, B., H.-J. Klumpen, and H. Gurney, *Dose calculation of anticancer drugs.* Expert Opinion on Drug Metabolism and Toxicology, 2008. **4**: p. 1307-1319.
5. Watters, J.W., et al., *Genome-wide discovery of loci influencing chemotherapy cytotoxicity.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(32): p. 11809-11814.
6. Longley, D.B., D.P. Harkin, and P.G. Johnston, *5-fluorouracil: mechanisms of action and clinical strategies.* Nat Rev Cancer, 2003. **3**(5): p. 330-8.
7. Johnston, P.G., et al., *Thymidylate synthase gene and protein expression correlate and are associated with response to 5-fluorouracil in human colorectal and gastric tumors.* Cancer Res, 1995. **55**(7): p. 1407-12.
8. Pullarkat, S.T., et al., *Thymidylate synthase gene polymorphism determines response and toxicity of 5-FU chemotherapy.* Pharmacogenomics J, 2001. **1**(1): p. 65-70.
9. Diasio, R.B., T.L. Beavers, and J.T. Carpenter, *Familial deficiency of dihydropyrimidine dehydrogenase. Biochemical basis for familial pyrimidinemia and severe 5-fluorouracil-induced toxicity.* J Clin Invest, 1988. **81**(1): p. 47-51.
10. Raida, M., et al., *Prevalence of a common point mutation in the dihydropyrimidine dehydrogenase (DPD) gene within the 5'-splice donor site of intron 14 in patients with severe 5-fluorouracil (5-FU)- related toxicity compared with controls.* Clin Cancer Res, 2001. **7**(9): p. 2832-9.
11. Pommier, Y., *DNA Topoisomerase I Inhibitors: Chemistry, Biology, and Interfacial Inhibition.* Chemical Reviews, 2009. **109**(7): p. 2894-2902.
12. Innocenti, F., et al., *Genetic Variants in the UDP-glucuronosyltransferase 1A1 Gene Predict the Risk of Severe Neutropenia of Irinotecan.* J Clin Oncol, 2004. **22**(8): p. 1382-1388.
13. Pommier, Y., et al., *Topoisomerase I inhibitors: selectivity and cellular resistance.* Drug Resist Updat, 1999. **2**(5): p. 307-318.
14. Sirikantaramas, S., M. Yamazaki, and K. Saito, *Mutations in topoisomerase I as a self-resistance mechanism coevolved with the production of the anticancer alkaloid camptothecin in plants.* Proceedings of the National Academy of Sciences, 2008. **105**(18): p. 6782-6786.
15. Lebwohl, D. and R. Canetta, *Clinical development of platinum complexes in cancer therapy: an historical perspective and an update.* Eur J Cancer, 1998. **34**(10): p. 1522-34.
16. Kostova, I., *Platinum complexes as anticancer agents.* Recent Pat Anticancer Drug Discov, 2006. **1**(1): p. 1-22.
17. Martin, L.P., T.C. Hamilton, and R.J. Schilder, *Platinum resistance: the role of DNA repair pathways.* Clin Cancer Res, 2008. **14**(5): p. 1291-5.

18. Bosl, G.J. and R.J. Motzer, *Testicular germ-cell cancer.* N Engl J Med, 1997. **337**(4): p. 242-53.

19. Welsh, C., et al., *Reduced levels of XPA, ERCC1 and XPF DNA repair proteins in testis tumor cell lines.* Int J Cancer, 2004. **110**(3): p. 352-61.

20. Koberle, B., et al., *Defective repair of cisplatin-induced DNA damage caused by reduced XPA protein in testicular germ cell tumours.* Curr Biol, 1999. **9**(5): p. 273-6.

21. Stewart, D.J., *Mechanisms of resistance to cisplatin and carboplatin.* Critical Reviews in Oncology/Hematology, 2007. **63**(1): p. 12-31.

22. Tang, S.C., *Strategies to decrease taxanes toxicities in the adjuvant treatment of early breast cancer.* Cancer Invest, 2009. **27**(2): p. 206-14.

23. Wang, Y., et al., *Resistance to microtubule-stabilizing drugs involves two events: beta-tubulin mutation in one allele followed by loss of the second allele.* Cell Cycle, 2005. **4**(12): p. 1847-53.

24. McGrogan, B.T., et al., *Taxanes, microtubules and chemoresistant breast cancer.* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 2008. **1785**(2): p. 96-132.

25. Marsh, S., et al., *Pharmacogenetic Assessment of Toxicity and Outcome After Platinum Plus Taxane Chemotherapy in Ovarian Cancer: The Scottish Randomised Trial in Ovarian Cancer.* J Clin Oncol, 2007. **25**(29): p. 4528-4535.

26. Green, H., et al., *mdr-1 single nucleotide polymorphisms in ovarian cancer tissue: G2677T/A correlates with response to paclitaxel chemotherapy.* Clin Cancer Res, 2006. **12**(3 Pt 1): p. 854-9.

27. Yamaguchi, H., et al., *Genetic variation in ABCB1 influences paclitaxel pharmacokinetics in Japanese patients with ovarian cancer.* International Journal of Gynecological Cancer, 2006. **16**(3): p. 979-985.

28. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-251.

29. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.

30. Fearnhead, N.S., B. Winney, and W.F. Bodmer, *Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model.* Cell Cycle, 2005. **4**(4): p. 521-5.

31. Fearnhead, N.S., et al., *Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas.* Proc Natl Acad Sci U S A, 2004. **101**(45): p. 15992-7.

32. Hillier, L.W., et al., *Whole-genome sequencing and variant discovery in C. elegans.* Nat Methods, 2008. **5**(2): p. 183-8.

33. Ossowski, S., et al., *Sequencing of natural strains of Arabidopsis thaliana with short reads.* Genome Res, 2008. **18**(12): p. 2024-33.

34. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

35. Wang, J., et al., *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-5.

36. Ley, T.J., et al., *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nature, 2008. **456**(7218): p. 66-72.

37. Kiss, M.M., et al., *High-Throughput Quantitative Polymerase Chain Reaction in Picoliter Droplets.* Anal Chem, 2008. **80**(23): p. 8975-8981.
38. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
39. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics., 2009. **25**(16): p. 2078-2079.
40. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet. **43**(5): p. 491-498.
41. Zhang, G., et al., *Statistical power of association using the extreme discordant phenotype design.* Pharmacogenet Genomics, 2006. **16**(6): p. 401-13.
42. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome Res, 2001. **11**(5): p. 863-74.
43. Souglakos, J., et al., *Ribonucleotide reductase subunits M1 and M2 mRNA expression levels and clinical outcome of lung adenocarcinoma patients treated with docetaxel/gemcitabine.* Br J Cancer, 2008. **98**(10): p. 1710-5.
44. Kuehl, P., et al., *Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression.* Nat Genet, 2001. **27**(4): p. 383-91.
45. Goyette, P., et al., *Human methylenetetrahydrofolate reductase: isolation of cDNA mapping and mutation identification.* Nat Genet, 1994. **7**(4): p. 551.
46. Cleaver, J.E., et al., *A summary of mutations in the UV-sensitive disorders: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy.* Hum Mutat, 1999. **14**(1): p. 9-22.
47. Dolan, M.E., et al., *Heritability and linkage analysis of sensitivity to cisplatin-induced cytotoxicity.* Cancer Res, 2004. **64**(12): p. 4353-6.
48. Bleibel, W.K., et al., *Identification of genomic regions contributing to etoposide-induced cytotoxicity.* Hum Genet, 2009. **125**(2): p. 173-80.
49. Simon-Sanchez, J., et al., *Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals.* Hum Mol Genet, 2007. **16**(1): p. 1-14.
50. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.
51. Wilson, J.F., et al., *Population genetic structure of variable drug response.* Nat Genet, 2001. **29**(3): p. 265-9.
52. Fijal, B.A., et al., *Candidate-gene association analysis of response to risperidone in African-American and white patients with schizophrenia.* Pharmacogenomics J, 2009. **9**(5): p. 311-8.

## Chapter 4

## GWAS-guided solution-phase target capture and mapping of adult-onset deafness in Border collies

This chapter contains content from the following manuscript:

*(Under revision at PLoS Genetics)* **Variation in genes related to cochlear biology is strongly associated with adult-onset deafness in Border collies**

Jennifer S. Yokoyama*, Ernest T. Lam*, Alison L. Ruhe, Carolyn A. Erdman, Katy R. Robertson, Aubrey A. Webb, D. Collette Williams, Melanie L. Chang, Marjo K. Hytönen, Hannes Lohi, Steven P. Hamilton and Mark W. Neff

*Equally contribution

The domestic dog offers a unique opportunity to study complex disorders similar to those seen in humans, but within the context of the much simpler genetic backgrounds of pure breeds, which represent closed populations. We searched for genetic risk factors of adult-onset deafness in the Border collie, a breed of herding dog that relies on acute hearing to perceive and respond to commands while working. Adult-onset deafness in Border collies typically begins in early adulthood and is similar to age-related hearing loss in humans. We first used a genome-wide approach to identify a region associated with the phenotype. Then, we used solution-phase capture and next generation sequence to look for variants in the region and attempted to identify the causative variant(s).

### *4.1 Abstract*

The domestic dog suffers from hearing loss that can have a profound impact on working ability and quality of life. We have identified a type of adult-onset hearing loss in Border collies that appears to have a genetic component, with an earlier age of onset (3-5 years) than typically expected for aging dogs (8-10 years). Studying this complex trait within pure breeds of dog may greatly increase our ability to identify genomic regions associated with risk of hearing impairment in dogs and in humans. We performed a genome-wide association study (GWAS) to detect loci underlying adult-onset deafness in a sample of 20 affected and 28 control Border collies. We identified a region on canine chromosome 6 that demonstrates extended support for association surrounding SNP Chr6.25819273 (p-value = $1.09 \times 10^{-13}$). To further localize disease-associated variants, targeted next-generation sequencing (NGS) of one affected and two unaffected dogs was performed. Through additional validation based on targeted genotyping of additional cases (n=23 total) and controls (n=101 total) and an independent replication cohort of 16 cases and 265 controls, we identified variants in *USP31* that were strongly associated with adult-onset deafness in Border collies, suggesting the involvement of the NF-κB pathway. We found additional support for involvement of *RBBP6*, which is critical for cochlear development. These findings highlight the utility of GWAS-guided fine-mapping of genetic loci using targeted NGS to study hereditary disorders of the domestic dog that may be analogous to human disorders.

### *4.2 Introduction*

Age-related hearing loss (presbycusis) occurs in humans with a prevalence of about 40% in individuals older than 65 years of age. It is associated with communication difficulties, isolation, depression and possibly even dementia in the severely affected [1]. There are extensive genetic contributions to hearing variation [2], which has an estimated heritability of 35-55% (reviewed in [3]). Studies in humans have identified risk-conferring variants in both the mitochondrial [4, 5] and nuclear genome (reviewed in [3]). A recent genome-wide association study (GWAS) performed in an isolated Finnish population identified the gene *IQ motif containing GTPase activating protein 2 (IQGAP2)* as a novel risk locus for hearing loss [6]. Modest support was also shown for another previously identified GWAS candidate, *metabotropic glutamate receptor 7 (GRM7)* [7]. Overall, however, the breadth of genetic variation underlying this common disorder remains unclear.

The domestic dog offers a unique opportunity to explore the genetic backgrounds of naturally occurring disorders that are analogous to human diseases. Genomic studies are particularly informative when a disorder of interest demonstrates a simpler inheritance pattern in dogs than in humans, suggesting one or a few main risk alleles. Deterioration of hearing with age is normal in dogs and corresponds with physiological changes in critical systems in the ear, including reduced spinal ganglion neuronal density in the cochlea [8]. Shimada et al. [9] reported that dogs with hearing loss demonstrated the same four types of lesions found in humans (as described by Schuknecht & Gacek [10]): sensory, neural, strial and cochlear conductive lesions. Physiological measurements of hearing ability using brainstem auditory evoked response (BAER) demonstrate similar patterns in dogs and humans, with high- and mid-range frequencies being the most severely affected [11, 12]. Thus, age-related hearing loss may be similar in both clinical presentation and underlying pathology in humans and in dogs.

Across breeds, presbycusis is estimated to begin at 8-10 years, when deterioration is observed at all frequencies [11]. However, adult-onset deafness in Border collies often has an earlier onset (3-5 years) than deafness resulting from the physiological aging of hearing organs. Distinct from other breeds, the Border collie has been selected for over 100 years to perceive and respond to whistle commands while working at distances of up to or over 800 meters from a handler. Being able to detect slight differences in whistle tones is essential to the function of a working Border collie, and even moderate hearing loss in one ear can have a major impact on working ability. Although relatively uncommon in Border collies, adult-onset deafness is considered especially problematic because hearing is so integral to the tasks for which these dogs are selectively bred and used. In addition, dogs afflicted by adult-onset deafness are often in their prime working years, with the average age of top working dogs around 7 years.

The earlier age of onset in affected Border collies suggests that adult-onset deafness is genetically influenced and possibly more severe than that observed in other breeds of dogs. Many of the affected dogs included in this study were reported by their owners to have one or more first-degree family members with similar deafness. We undertook a study to identify genetic risk factors and address concerns regarding adult-onset deafness among Border collies, as well as potentially gain information about analogous human conditions.

### *4.3 Methods*

### ***Data collection***

*Ethics Statement*. All work related to animals was performed with the approval of the Institutional Animal Care and Use Program at the University of California, San Francisco (AN079848-02). Collection of blood samples in Finland was approved by the Animal Ethics Committee at the State Provincial Office of Southern Finland (ESLH-2009-07827/Ym-23). The canine samples used were provided by private dog owners, who consented to the use of de-identified data for research purposes.

*Samples*. Whole blood samples (3-8 mL) from a total of 48 purebred Border collies collected in the United States (U.S.) were used for primary GWAS. Samples from 20 affected working Border collies recruited among owners from the sheepdog/herding community were collected specifically for this genetic survey of risk loci for adult-onset deafness. Twenty-eight control samples (unrelated at the grandparental level, per pedigree analysis) were collected at sheepdog trials or sent directly to the laboratory by owners and breeders in the context of ongoing genetic studies of canine behavior and complex disease. The 20 adult-onset deafness cases included 9 males and 11 females, and the 28 controls included 15 males and 13 females (mean age of controls, 6.6 years). One of the cases and two controls were also sequenced using next-generation sequencing (NGS) technology. An additional 14 U.S. controls and 3 cases and 59 controls collected in Finland were used for follow-up genotyping of candidate variants. Finally, samples from 16 cases and 265 controls were also collected in the U.S. to serve as an independent replication cohort. All follow-up and replication samples were from purebred Border collies. Although consisting primarily of distinct breeding lines, the Finnish dogs demonstrated similar allele frequencies for the genotyped variants as the U.S. dogs, and thus both groups were analyzed together. The use of a covariate to account for difference in country of origin/breeding line did not change the results of association analysis. DNA was extracted using standard protocols. A summary of the samples is provided in Table 4.1.

*Phenotypes*. Adult-onset deafness phenotypes were assigned based on owner responses to verbal questions to determine whether or not a sampled dog exhibited hearing loss that had developed in adulthood (i.e., deafness that was not congenital). Hearing loss was determined indirectly by owner observations of working dogs that were previously responsive to verbal and whistle commands given in both home and working conditions, but as adults demonstrated significant decreases in response or apparent inability to hear commands. Such loss of hearing was often observed to take place over the course of several months or years. Some owners said that they did not notice any significant changes in their dogs' hearing ability until much later in the dogs' lives, but they suspected that the dog was "compensating" in the work environment by observing the handlers' physical cues or by moving closer to the handler when commands were being given. Controls for the primary GWAS portion of the study were herding Border collies that met two criteria: 1) genetic clustering in the same group as affected dogs (genetic matching), and 2) no hearing loss indicated in the health sections of behavioral questionnaires completed by owners at the time of sample collection. For follow-up and replication genotyping, U.S. cases were identified as above and controls were defined as dogs that displayed no hearing loss as indicated by owners at the time of sample collection. For all Finnish dogs, deafness phenotypes were obtained through owner interviews via questionnaires.

*Genome-wide genotyping*. SNP genotyping was performed on the Affymetrix Custom Canine Array v2.0 according to the manufacturer's protocol, a perfect-match-only array targeting 127,000 SNPs (Affymetrix). Genotypes were called using the BRLMM-P algorithm in Affymetrix Power Tools (apt-1.12.0). Genotype quality control (QC) was first implemented for all of the samples we genotyped on the Affymetrix array for ongoing studies of complex disease (n = 275), which included unrelated dogs as well as a subset of related dogs to assess Mendelian errors. SNP exclusion criteria for the full set were call rates by marker and by individual < 95%; concordance of replicate control sample genotypes across all genotyping runs < 100%; X-chromosome markers; deviations from Hardy-Weinberg equilibrium with p-values < 0.001; minor allele frequency < 0.02; and Mendelian errors > 5% per SNP and > 10% per family. This filtering resulted in a primary dataset of about 40,000 SNPs. After additional QC on only the 48 unrelated samples included in the GWAS for adult-onset deafness (exclusions: SNP call rate < 95%, MAF < 0.05), approximately 30,000 SNPs were retained for final analysis. QC was performed using Stata10/MP (StataCorp LP) and PLINK (v1.06-1.07 [13]).

*Target capture and next-generation sequencing*. Genomic library sample preparation was performed using the Illumina single-end library sample preparation kit. Sample preparation was carried out according to the manufacturer's instructions, except for slight modifications as follows: 3 µg of genomic DNA were sheared via sonication (S-4000 with 2.5" diameter cup horn, Misonix, Inc.); all purification steps were performed using Agencourt AMPure XP magnetic beads (Beckman Coulter, Inc.); seven cycles of ligation-mediated PCR were used for library amplification. Sample libraries were run on a Bioanalyzer 2100 for DNA quantitation and confirmation of fragment size distribution (High Sensitivity DNA Kit, Agilent Technologies). For targeted sequencing, we performed solution-based capture with the Agilent SureSelect Target Enrichment System Kit. Briefly, a custom panel of 120-base cRNA oligos was designed to target 1000 bp upstream and downstream of 73 predicted gene sequences based on mammalian alignments (or in one case, frog) to CanFam2 in the candidate region on CFA6 (Table 4.2). The target regions were covered by approximately 43,000 probes that were designed for 3X coverage. The prepared genomic libraries were hybridized to the panel of biotin-labeled "bait" oligos for 24 hours. Targets were pulled down via streptavidin magnetic beads, purified, and enriched through 13 cycles of PCR amplification. Samples were single-end sequenced on an Illumina Genome Analyzer IIx for 76 cycles.

*Dye-terminator sequencing*. Three variants were selected for genotyping via dye-terminator sequencing. All samples included in the study were sequenced. A PCR amplicon was designed for each region, and sequencing was performed in the forward and reverse directions (primer sequences provided in Table 4.3). We used 1 µL of 1 ng/µL of DNA as input for each standard PCR reaction. Platinum-Taq polymerase was used to amplify segments with a 58°C touchdown protocol in the presence of 0.4 μM primer, 100 μM dNTPs, 2.5 mM Mg and 1 mM betaine.

***Data analysis***

*Genome wide association study.* Primary GWAS analysis was performed using the beta version of Efficient Mixed-Model Association eXpedited (EMMAX) [14]. We used a mixed model-based analysis to account for population stratification or cryptic relatedness that may have been present in the sample (as pedigrees were not available for cases). In addition, allelic associations with one million permutations were performed in PLINK to further assess association strength and rule out false positives. Permutation analysis consists of reassigning case-control labels randomly and identifying the distribution of possible associations at each SNP for all random case-control assignments. Haplotype analysis was performed in PLINK, with results visualized using the Genome Variation Server (GVS) (http://gvs.gs.washington.edu/GVS/index.jsp) and Haploview.

*Next generation sequencing.* Bowtie [15] was used for read alignment against CanFam2, allowing up to 2 mismatches in the first 60 bases of a given read. SAMtools [16], Picard (http://picard.sourceforge.net/), BEDTools [17], and the Genome Analysis Toolkit (GATK, [18]) were used for post-alignment processing. Multi-sample realignment around potential insertion/deletions (indels) and base quality score recalibration were both performed prior to variant calling by GATK's Unified Genotyper. Indels were called using Dindel [19]. ANNOVAR [20] was used to annotate and prioritize variants.

Phastcons4way scores, which provide a measure of conservation based on multi-species alignment, were obtained from the UCSC Genome Browser [21].

*Dye-terminator sequencing.* Genotype calls for the three variants were made manually by inspection of sequence traces. Association testing was performed in PLINK, and false positives were assessed by subsequent permutation testing. Allele frequencies of variants between the Border collies from the U.S. and Finland were not significantly different, according to Fisher's exact test for homogeneity; further, all associations remained when ancestry was added as a covariate (data not shown). The two groups were thus treated as a single group in follow-up analysis.

*Meta-analysis of candidate variants.* A sample size weighted analysis based on p-values generated in the primary and replication cohorts was performed using METAL [22].

## *4.4 Results*

### *Adult-onset deafness in Border collies*

The exact age of onset of hearing deterioration is often difficult to ascertain in pet dogs, because subtle changes may go unnoticed by pet owners and because dogs are known to compensate for hearing loss [23]. However, the owners of the dogs included in this study estimated the age of onset of hearing loss based upon close observations of behavioral characteristics in working dogs indicating poor hearing (e.g., reduced call distance, poor performance). The average estimated age of onset was 4.3 years (standard error of 0.5 years), with a range of 1-9 years.

### *Genome-wide association study*

A total of 48 unrelated Border collies (20 cases, 28 controls) were used for the primary association study (Table 4.1). Following quality control of the genotype data, 30,231 SNPs were retained for genetic mapping. Genome-wide association analyses with EMMAX identified a region on CFA6, at approximately 25 Mb, with strong regional support demonstrated by neighboring SNPs with significant p-values (Figure 4.1). In total, 25 markers exhibited significance beyond a Bonferroni-corrected threshold ($p = 1.65 \times 10^{-6}$ for 30,231 tests). The strongest finding was an intergenic SNP, Chr6.25819273, with a p-value of $1.09 \times 10^{-13}$ (Table 4.4). The closest predicted gene to this SNP is *HS3ST2*, approximately 24 kb downstream of the marker. HS3ST2 is a member of the heparan sulfate biosynthetic enzyme family, and is expressed predominantly in the brain [24].

Associations were also assessed through permutation analysis in PLINK. One million permutations yielded genome-wide permutated p-values that achieved genome-wide significance (Table 4.4). Analyses of copy number variation using genome-wide SNP data did not reveal evidence of structural changes associated with hearing loss. Association modeling suggested an autosomal recessive mode of inheritance for adult-onset deafness in Border collies.

### *Fine-mapping*

The large candidate region identified on CFA 6 was syntenic with human 16p12.1-p12.3, which encompasses the human autosomal recessive deafness locus *DFNB22* [25]. A candidate of immediate interest within this region was the gene *OTOA*, defects of which were implicated in a case of prelingual sensorineural deafness in a consanguineous Palestinian family [25]. We performed PCR amplifications of the 28 exons and a highly conserved non-coding region. PCR products were sequenced and analyzed for mutations in affected dogs. None of the observed polymorphisms tracked specifically in affected dogs.

Given the large region of association and lack of polymorphisms in the strong candidate gene *OTOA*, we next narrowed the critical region for the 25-Mb locus by haplotype analysis. We detected a 7-SNP haplotype that was homozygous in all cases and in only one control sample (see *Discussion*), as well as a larger 11-SNP haplotype that was found in 19 of 20 cases (and was present in the same control, Figure 4.2). For sequencing via target capture and next-generation sequencing (NGS), we selected an affected dog that was homozygous for the extended 11-SNP risk haplotype at 25 Mb (Figure 4.2). Two dogs that did not carry the candidate risk haplotype were also sequenced after target capture by NGS.

The extended risk haplotype spanning 25.5-25.9 Mb was used to guide mutation discovery with the NGS data. We identified predicted genes based on synteny and designed a solution-based target capture mixture to target exons and introns, along with at least 1 kb of upstream and downstream untranslated regions. This capture design encompassed 2.3 Mb of sequence and included 73 genes at 25 Mb annotated in the canine or other genomes (Table 4.2). NGS rendered over 30 million reads per sample (Table

4.5), and over 90% of the reads from each sample could be aligned to the dog sequence (CanFam2). About 75% of the total targeted sequence had >10X coverage, and nearly 70% had >30X coverage (Table 4.6).

The numbers and types of variants identified are summarized in Table 4.7. One strong non-synonymous SNP (nsSNP) candidate, Chr6.25714052, is located in exon 17 of *USP31*, which encodes an ubiquitin specific peptidase. It is an A>G variant that is predicted to cause an I847V change in the resulting protein product. The position is highly conserved, with a phastCons score of 0.95 (Table 4.8), although SIFT predicted the change to be tolerated (SIFT score of 0.66). Also of note in *USP31* is an intronic T>G SNP (Chr6.25681850) that is very highly conserved (phastCons score of 0.98) and is 5 bp away from an intron-exon boundary. This variant was called G/G in the case and T/T in both controls. Both variants are located within the risk haplotype.

Another candidate nsSNP, Chr6.24500625, is in exon 18 in *RBBP6*, encoding a retinoblastoma binding protein. This nsSNP changes threonine to asparagine at residue 1,397. This G>T variant was called T/T in the case and G/G in both controls. SIFT predicted the change to be tolerated (SIFT score = 0.69; Table 4.8). Although the conservation score for this SNP is low (phastCons = 0.001) and the variant is located upstream of the main risk haplotype, RBBP6 (also known as PACT) plays a critical role in ear development and hearing; disruption of the gene has been shown to cause congenital hearing impairment in mice [26] and suggests high relevance to hearing loss in dogs. The sequencing data from 25 Mb did not exhibit variants in *OTOA* that were homozygous in the case but not in controls. Therefore, we did not consider this to be the causative gene. Although small insertions/deletions (indels) were found in the mapped intervals, none of these variants appeared to be causal.

The three variants described above, Chr6.24500625 in *RBBP6*, Chr6.25681850 and Chr6.25714052 in *USP31*, were the most compelling for follow-up genotyping analyses due to biological implications (*RBBP6*) and location within the risk haplotype (*USP31*), and were analyzed both for validation (primary mapping cohort) and replication (independent cases and controls).

### *Validation*

Genotyping was performed via dye-terminator sequencing for the three chosen variants. All three showed associations with adult-onset deafness (Table 4.9). For replication analysis, we genotyped an independent Border collie cohort of 16 cases and 265 controls. All three SNPs were strongly associated with adult-onset deafness (Table 4.9), replicating our previous mapping results. Meta-analysis of the combined primary and replication cohorts yielded even stronger associations for all three variants. The strongest association was found for the potential splice variant of *USP31*, Chr6.25681850, with $p = 6.16 \times 10^{-22}$ (Table 4.9).

## *4.5 Discussion*

Our results represent the first GWAS of adult-onset deafness in the domestic dog. We demonstrated the successful application of target capture for next-generation sequencing (NGS) in the dog. The region implicated by GWAS in our study is syntenic to regions implicated in congenital sensorineural deafness in humans.

In this study, we identified three strong candidate coding and non-coding variants associated with adult-onset deafness. The strongest is Chr6.25681850, an intronic SNP in *USP31* that is 5 bp from an intron-exon boundary and may play a role in alternate splicing (as annotated in humans). Preliminary studies of mRNA collected from peripheral blood samples from two dogs harboring this variant did not suggest changes in RNA splicing in this region, though tissue-specific changes in RNA regulation cannot be ruled out. *USP31* is an ubiquitin-related gene that has been linked to Parkinson's disease in humans [27]. The implication of an ubiquitin-related gene in adult-onset deafness is particularly intriguing given the histological findings of Shimada et al. [9], which included ubiquitin-positive granules in neuropils of the cochlear nuclei of aging dogs. USP31 has also been shown to regulate NF-κB activation; NF-κB deficiency is associated with increased levels of cochlear apoptosis and hearing loss [28, 29]. Despite its location outside the main risk haplotype implicated in the primary GWAS, the second strongest association was the nsSNP Chr6.24500625, which is exonic to *RBBP6*, a gene previously implicated in hearing in a knockout mouse model [26]. In addition to roles in development, RBBP6 may also be involved in chaperone-mediated ubiquitination and protein quality control [30], suggesting another potential role in pathology. A second *USP31* SNP, Chr6.25714052, was also associated with adult-onset deafness in our cohort, although this locus had the lowest odds ratio of the three candidate loci.

There are several caveats to the present study. A recent human GWAS for presbycusis adjusted phenotypes for hearing thresholds according to age and sex, due to observed differences in hearing threshold variability in males and females [6]. We elected not to correct for sex in our canine study because such sexual dimorphism is not yet established in aging dogs [9, 11, 12, 23]. Further, we did not adjust for age because the age of onset for our sample cohort, which is likely a specific trait of this form of hearing loss, was owner-estimated. The mean age of our control group was 6.6 years, which is close to the range of hearing loss onset. Therefore, it is possible that dogs categorized as "controls" may, at later stages in life, demonstrate hearing loss similar to that observed in cases. For example, one interesting case involves a dog that was classified as a control at the time of collection (41 months old), but was later shown to carry the 11-SNP risk haplotype we identified in affected dogs (this dog is indicated by the asterisk in Figure 4.2). This dog was later found to have several deaf siblings. In the follow-up SNP genotyping cohort, several Finnish dogs classified as controls by owner questionnaires were also found to carry one or more of the risk alleles identified during NGS (Table 4.9). Two of these dogs were later found to have had changes in hearing since initial sample collection, and further inquiry uncovered additional family histories of hearing loss in both dogs' pedigrees. However, the misclassification of cases as controls would only reduce analytical power to detect genetic associations, and would not result in spurious associations. Given the strengths of the associations we identified on CFA6, this does not seem to be a concern. Similarly, the presence of the risk haplotype in the homozygous state in all cases suggests that we are not detecting phenotypic heterogeneity influenced by another locus, such as occult congenital unilateral pigmentation-related forms of deafness.

Another caveat stems from the fact that we performed target enrichment for selected regions (i.e., all predicted genes) of extended association loci, and therefore non-coding variants far outside of known or predicted genes were potentially missed. Target enrichment results in uneven coverage, so variants may be missed because not all positions are covered equally well, although the reference sequence in the regions being captured appear to be well assembled (Figures 4.3-4.4, Table 4.6). Finally, the magnitude of our findings on CFA6 in the primary GWAS likely overshadowed signals from other regions, even if modifying loci were present.

A major strength to be highlighted was the ability to obtain accurate phenotype information for adult-onset deafness strictly from owner observation, as phenotypes were not based on BAER testing or clinical diagnosis. Although a subset of phenotypes were obtained by questionnaires (in Finland), the majority of samples – including all of the samples used for the primary GWAS – were phenotyped based strictly on owner comments regarding recognition of hearing changes in adult dogs (see Materials and Methods for more details). The strength of the primary association findings on CFA6 for adult-onset deafness is a testament to the sheepdog handlers' ability to identify changes in their dogs' behavior that were reflective of hearing loss. Similar to a parent's intimate knowledge of his or her child's behavior, most phenotyping in community-based dog samples relies critically on owner ability to reliably recount behavior in different contexts. For researchers interested in the biological underpinnings of behavior, our findings strongly suggest that observations of working dogs by their handlers – with or without the use of formal questionnaires – may be specific enough for phenotyping in genetic studies given their close working relationship.

Although we observed robust associations and replications, none of the candidate SNPs we identified tracked perfectly with adult-onset deafness. This discrepancy has several possible explanations: 1) adult-onset deafness in the Border collie is a multigenic trait, 2) the risk locus shows incomplete penetrance, or 3) the variants we identified are in linkage disequilibrium with the true disease-causing mutation. The fact that the *RBBP6* SNP demonstrated a stronger association than the second *USP31* SNP, Chr6.25714052, likely reflects extended linkage in cases that was not readily apparent in haplotype analyses, and may provide information regarding the location of the true causative variant. Given that the 7-SNP homozygous haplotype is present in all cases, it is likely that the variants we identified, which do not track perfectly with larger samples of cases, are more recent in origin than the common tagging SNPs utilized in array genotyping. This would suggest that the causative variant has occurred within the context of a broader, ancestral haplotype. The causative mutation for adult-onset deafness may be a non-coding variant between Chr6.24500625 and Chr6.25681850 that was not captured during target enrichment, and structural variation may also be missed with this technology. Numerous mapping studies in the dog have identified structural variants as causative mutations of traits or disorders [31].

The risk allele of the most strongly-associated SNP from NGS exhibited a frequency of 0.23-0.31 in our Border collie control sample (Table 4.9). Future studies may clarify whether this risk allele occurs at similar frequencies in other breeds of dog. Alternative mapping strategies utilizing highly polymorphic microsatellite markers in haplotypes and including different breeds of dog may allow for more refined mapping of structural variants underlying adult-onset deafness. In light of our strong genetic findings, longitudinal studies of dogs that carry risk alleles are warranted for further phenotypic characterization, including histopathologic examination of the middle ears and cochlea. Such investigations may allow us to further characterize and explore the hypothesis that these animals are affected by pure sensorineural deafness, as demonstrated by BAER testing. Observations of the effects of risk variants on aspects of hearing throughout the aging process could provide critical prognostic information for the development of diagnostic or therapeutic tools for use in clinical contexts in both dogs and in humans. It is possible that hearing loss is identified earlier by handlers of dogs for which working ability depends strongly on hearing acuity, such as working Border collies. Physiological findings may thus be particularly relevant to studies of other utility-bred dogs, in addition to studies of hearing loss that naturally occurs in geriatric dogs.

In conclusion, we identified candidate variants on CFA6 that are strongly associated with adult-onset deafness in Border collies, with promising implications for future pre-morbid identification of at-risk dogs or applications to human studies. Preliminary causative variant fine-mapping analyses indicate that variants in *USP31* and *RBBP6* may be involved in disease etiology. Future studies to elucidate the roles of these variants in canine adult-onset hearing loss will include haplotype mapping for the detection of structural variations and longitudinal studies of gene effects on hearing electrophysiology trajectories and outcomes.

### *4.6 Funding sources*

This research is supported in part by NIH award to S.P.H. (MH084149). J.S.Y was supported in part by NIH-NIMH Predoctoral Training Consortium in Affective Science (T32 MH20006). J.S.Y and E.T.L were supported in part by NIH Training Grant T32 GM007175. Part of this work is supported by intramural funds from both the Veterinary Genetics Laboratory at UC Davis and the Van Andel Research Institute to M.W.N.

### *4.7 References*

1. Gates, G.A. and J.H. Mills, *Presbycusis.* Lancet, 2005. **366**(9491): p. 1111-1120.
2. Karlsson, K.K., J.R. Harris, and M. Svartengren, *Description and primary results from an audiometric study of male twins.* Ear Hear., 1997. **18**(2): p. 114-120.
3. Liu, X.Z. and D. Yan, *Ageing and hearing loss.* J Pathol., 2007. **211**(2): p. 188-197.
4. Bai, U., et al., *Mitochondrial DNA deletions associated with aging and possibly presbycusis: a human archival temporal bone study.* Am J Otol., 1997. **18**(4): p. 449-453.
5. Fischel-Ghodsian, N., et al., *Temporal bone analysis of patients with presbycusis reveals high frequency of mitochondrial mutations.* Hear.Res, 1997. **110**(1-2): p. 147-154.
6. Van Laer, L., et al., *A genome-wide association study for age-related hearing impairment in the Saami.* Eur J Hum Genet, 2010.
7. Friedman, R.A., et al., *GRM7 variants confer susceptibility to age-related hearing impairment.* Human Molecular Genetics, 2009. **18**(4): p. 785-796.
8. Knowles, K., et al., *Reduction of spiral ganglion neurons in the aging canine with hearing loss.* Zentralbl.Veterinarmed.A, 1989. **36**(3): p. 188-199.
9. Shimada, A., et al., *Age-Related Changes in the Cochlea and Cochlear Nuclei of Dogs.* The Journal of Veterinary Medical Science, 1998. **60**(1): p. 41-48.
10. Schuknecht, H.F. and M.R. Gacek, *Cochlear pathology in presbycusis.* Ann.Otol.Rhinol.Laryngol., 1993. **102**(1 Pt 2): p. 1-16.
11. Ter, H.G., et al., *Effects of aging on brainstem responses to toneburst auditory stimuli: a cross-sectional and longitudinal study in dogs.* J Vet.Intern Med, 2008. **22**(4): p. 937-945.
12. Ter, H.G., et al., *Effects of aging on inner ear morphology in dogs in relation to brainstem responses to toneburst auditory stimuli.* J Vet.Intern Med, 2009. **23**(3): p. 536-543.
13. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-575.
14. Kang, H.M., et al., *Variance component model to account for sample structure in genome-wide association studies.* Nat Genet, 2010. **42**(4): p. 348-354.
15. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
16. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics., 2009. **25**(16): p. 2078-2079.
17. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics., 2010. **26**(6): p. 841-842.
18. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-1303.
19. Albers, C.A., et al., *Dindel: accurate indel calls from short-read data.* Genome Res, 2011. **21**(6): p. 961-973.

20. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucl.Acids.Res., 2010. **38**(16): p. e164-e164.
21. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Res., 2002. **12**(6): p. 996-1006.
22. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics., 2010. **26**(17): p. 2190-2191.
23. Strain, G.M., *Aetiology, prevalence and diagnosis of deafness in dogs and cats.* Br.Vet.J, 1996. **152**(1): p. 17-36.
24. Shworak, N.W., et al., *Multiple isoforms of heparan sulfate D-glucosaminyl 3-O-sulfotransferase. Isolation, characterization, and expression of human cdnas and identification of distinct genomic loci.* J Biol Chem., 1999. **274**(8): p. 5170-5184.
25. Zwaenepoel, I., et al., *Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying acellular gels, is defective in autosomal recessive deafness DFNB22.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(9): p. 6240-6245.
26. Rowe, T.M., et al., *A role of the double-stranded RNA-binding protein PACT in mouse ear development and hearing.* Proc.Natl.Acad.Sci.U.S.A, 2006. **103**(15): p. 5823-5828.
27. Lockhart, P.J., et al., *Identification of the human ubiquitin specific protease 31 (USP31) gene: structure, sequence and expression analysis.* DNA Seq., 2004. **15**(1): p. 9-14.
28. Lang, H., et al., *Nuclear factor kappaB deficiency is associated with auditory nerve degeneration and increased noise-induced hearing loss.* J Neurosci, 2006. **26**(13): p. 3541-3550.
29. Tzimas, C., et al., *Human ubiquitin specific protease 31 is a deubiquitinating enzyme implicated in activation of nuclear factor-kappaB.* Cell Signal., 2006. **18**(1): p. 83-92.
30. Kappo, M.A., et al., *Solution structure of RING finger-like domain of retinoblastoma-binding protein-6 (RBBP6) suggests it functions as a U-box.* J Biol Chem, 2012. **287**(10): p. 7146-58.
31. Alvarez, C.E. and J.M. Akey, *Copy number variation in the domestic dog.* Mamm.Genome, 2012. **23**(1-2): p. 144-163.

## Chaper 5

## Nanomapping of the MHC region: structural variation analysis and *de novo* sequence assembly

This chapter contains content from the following manuscript:

*(In press at Nature Biotechnology)* **Nano-mapping for structural variation analysis and sequence assembly.**

Ernest T. Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, and Pui-Yan Kwok

### *5.1 Abstract*

We describe nano-mapping, a method suitable for facilitating *de novo* assembly of next-generation sequencing (NGS) data, haplotype and structural variation analysis, and comparative genomic analysis. Using a nanofluidic device containing thousands of nanochannels, we produced highly accurate, haplotype-resolved, sequence motif maps hundreds of kilobases in length by sequence-specific fluorescent labeling and automated high resolution single-molecule imaging of DNA molecules uniformly stretched in silicon-based nanochannels. We demonstrated the utility of our approach by constructing sequence motif maps of the human major histocompatibility complex (MHC) region with bacterial artificial chromosome (BAC) clones covering the region from two BAC libraries. Ninety-five previously sequenced BACs from the PGF and COX libraries were analyzed as individual BACs and as mixtures. We also sequenced the clones and assembled the sequencing data obtained. Our results show that sequence motif maps

provide useful scaffolds for *de novo* assembly of sequencing data generated from structurally complex regions in diploid organisms and for characterization of structural variants.

### *5.2 Introduction*

Despite recent advances in base-calling accuracy and read length, *de novo* genome assembly and structural variant analysis using “short read” shotgun sequencing remain challenging. Most resequencing projects rely on mapping the sequencing data to the reference sequence to identify variants of interest [1]. When whole genome assembly is attempted, it is done by paired-end sequencing to provide scaffolds for assembly [2]. As cloning of large DNA fragments is difficult, small insert libraries of varying sizes may be prepared for paired-end sequencing, thus limiting the resolution of haplotypes and increasing the complexity, time, and cost of the sequencing project. In addition, complex genomic loci, such as the major histocompatibility (MHC) region, important for infectious and autoimmune diseases [3], contain highly repetitive sequences and are particularly challenging for sequence assembly. Robust technologies that can aid in *de novo* sequence assembly are therefore sorely needed as whole genome sequencing becomes more widely adopted.

Emerging whole genome scanning techniques have revealed the prevalence and importance of structural variation. Detecting copy number variation often relies on detection of relative signal intensities by array-based or quantitative PCR-based technologies. Array-based methods, such as array-based comparative genomic hybridization (aCGH) have been used extensively in interrogation of copy number variation in the human genome [4, 5]. However, except for deletions, these methods do not provide positional information regarding the locations of copy number variants (CNVs) and cannot detect balanced structural variation, such as inversions or translocations [6]. Paired-end mapping techniques, traditionally by Sanger sequencing and now by next-generation sequencing [7] generally have low sensitivity in repetitive regions, where much structural variation lies [8]. Recent efforts to characterize CNVs in human genomes at high resolution involved paired-end mapping of clones, but this approach, while useful for exploratory studies in this small sample set, is too labor-intensive and time-consuming to be applicable for analysis of large numbers of individuals, and the resolution was no better than 8 kb [9].

Restriction mapping was instrumental in the Human Genome Project. One approach to address drawbacks of traditional restriction mapping is optical mapping [10]. In this approach, large DNA fragments are stretched and immobilized on glass slides and cut *in situ* with restriction enzymes. Optical mapping was used to construct ordered restriction maps for whole genomes [11-14], and it provided scaffolds for shotgun sequence assembly and validation [15, 16]. However, this method is limited by its low throughput, non-uniform DNA stretching, imprecise DNA length measurement, and high error rates.

Here we report a new, highly accurate, high-throughput nano-mapping technique that is being optimized for general use. The core technology is a commercially available nanofluidic chip that contains nanochannels that keep long DNA molecules in a consistent, elongated state. Fluorescently labeled DNA molecules are streamed into the nanochannels, held still, and imaged automatically on the multi-color NanoAnalyzer®

1000 instrument (BioNano Genomics, La Jolla, CA). After imaging, additional sets of DNA molecules are streamed into the nanochannels for imaging. This process is repeated many times until the DNA is depleted or the nanochannels are rendered unusable due to clogging.

Each nanofluidic chip contains three devices. Each nanochannel device consists of ~4000 channels that are 0.4 mm in length and 45 nm in diameter. Using 193 nm lithography in a nanofabrication process on the surface of a silicon substrate, nanochannel array chips are produced with precise diameters. Because DNA has a persistence length of ~50 nm, DNA molecules in the 45 mm nanochannels cannot fold back on themselves and are forced by physical confinement to be in the elongated, linearized state [17, 18]. As long DNA molecules in solution exist as coiled balls, a gradient region consisting of pillars and wider channels is placed in front of the nanochannels to allow the DNA molecules to uncoil as they flow toward the array (Figure 5.1) [19]. In this region, the physical confinement is sufficiently dense that the molecules are forced to interact with the pillars, yet sufficiently sparse that the DNA is free to uncoil. Once uncoiled, the DNA can then be efficiently flowed into the array in a linear manner.

Our nano-mapping approach combines robust sequence specific labeling, consistent linearization of extremely long DNA molecules in nanochannel arrays, automated imaging, high resolution single-molecule size measurements, and map construction. It provides a simple technique for mapping complex regions or whole genomes, and facilitates sequence assembly with long-range scaffolding information and structural variation analysis.

### *5.3 Methods*

*Sample preparation and data collection*. We obtained BAC clones in LB slabs from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute (http://bacpac.chori.org/) from the BAC libraries CHORI-501 and CHORI-502. All DNA samples used in the study were prepared using Qiagen's Large-Construct Kit. To prepare BAC mixtures, we grew 8 mL cultures of each BAC in LB containing 20 ug/mL chloramphenicol overnight and combined the separate cultures before proceeding with DNA extraction of the BACs as a pool. The DNA samples were quantified using NanoDrop 1000 (Thermal Fisher Scientific) and their quality assessed using pulsed-field gel electrophoresis. One milligram BAC DNA was linearized with 2 U of NotI or BsiWI and nicked with 0.5 U nicking endonuclease Nt.BspQI (New England BioLabs, NEB) at 37 $^{o}$C for 2 hours in NEB Buffer 3. [Note: NotI and BsiWI both cut in the vector but also cut several times within the clone insert. Linearizing with the two enzymes in separate reactions produces overlapping DNA fragments and minimizes the number of gaps in the final map.] The resultant DNA fragments were labeled with 25 nM Alexa546-dUTP (Invitrogen) and Vent (exo-) (NEB) for 1 hour at 72 $^{o}$C. The backbone of above fluorescently tagged DNA (5 ng/uL) was stained with YOYO-1 (3 nM; Invitrogen). DNA was loaded in BioNano Genomics nanochannel arrays by electrophoresis of DNA. First, twelve volts were applied to concentrate the DNA at the entrance of the channels. Thirty volts were applied to move DNA into the nanochannels, and 10 V was applied to distribute the DNA in the nanochannels. Linearized DNA molecules were imaged using blue and green lasers for YOYO-1 and Alexa546 on the BioNano Genomics NanoAnalyzer automated imaging system.

*Sequence motif map generation*. The DNA molecule (YOYO-1) and locations of fluorescent labels (Alexa546) along the length of each molecule were detected using the software package, NanoStudio. A set of label locations of each DNA molecule comprises the individual DNA molecular map. To take into account sizing errors and differences in the length of molecules, we transformed the data using a sliding window calculation. Each sequence motif map was thus made to have the same dimensions. We performed complete pairwise comparison of all single-molecule sequence motif maps and built a Euclidean distance matrix of the sequence motif maps. We then used the matrix as input for unsupervised hierarchical clustering of individual sequence motif maps based on Ward's method using the R package *fastcluster*. After unsupervised clustering, individual sequence motif maps were grouped and analyzed together. Peaks corresponding to signal coming from true nick sites were fit and called based on a Gaussian model. Clusters corresponding to different orientations of otherwise the same BAC are combined. False signal were filtered out based on a set threshold. The consensus clusters were then joined based on overlap to form the sequence motif map for the MHC region [20].

*Next-generation sequencing of BAC clones*. We pooled and performed paired-end 100bp sequencing of the two sets of BACs in one HiSeq 2000 lane each. We obtained 131 million and 142 million reads for PGF and COX, respectively. We first aligned the reads to the current reference genome build hg19 using Bowtie [21] allowing only unique, properly paired alignments and for up to 2 mismatches in the alignment. Duplicate reads with identical end-coordinates are removed using Picard (http://picard.sourceforge.net/). We then performed *de novo* assembly of the reads using SOAPdenovo [22]. Multiple k-mer lengths were tested to maximize assembly contiguity, and we chose ones that maximized the assembly N50. The assembly was refined by using GapCloser to further bridge gaps in the contigs. We also ran Velvet with different k-mer lengths for *de novo* assembly for the same dataset. “LONGSEQUENCES” was enabled to optimize assembly of long contigs [23]. Discussion of assembly throughout the manuscript is based on results from SOAPdenovo since the contigs from Velvet were much shorter.

### *5.4 Results*

To demonstrate the utility of our approach, we used nano-mapping to construct sequence motif maps of 95 BAC clones covering the 4.7 Mb MHC region from two individuals (PDF and COX libraries used by the MHC Haplotype Consortium [24, 25]). Subsequently, we performed *de novo* sequence assembly using next generation sequencing reads. The sequence-motif maps and sequencing contigs were then compared to the reference sequences reported by the MHC Haplotype Consortium as confirmation and to potentially uncover differences.

#### *Generation of sequence motif maps by nano-mapping*

Nano-mapping consists of four steps. It begins with sequence-specific labeling followed by linearization of the labeled long DNA molecules, imaging, and map construction as illustrated in Figure 5.2 with a 183 kb BAC clone. A nicking endonuclease was used to introduce single-strand nicks in the dsDNA at specific sequence motifs. Fluorescent dye-conjugated nucleotides (Alexa 546 dUTP) were then incorporated at these sites by Vent (exo-) polymerase (Figure. 5.2A) [26],[27]. The labeled DNA molecules were stained with the DNA intercalating dye, YOYO-1, loaded onto a nanochannel array chip, and introduced into the nanochannels (cross section 45 nm by 45 nm) by applying an electric field. The long coiled DNA molecules in free suspension were gradually moved through a series of micro- and nano-fluidic structures and stretched into linear form through entropic-confinement inside the nanochannels [19]. Once the nanochannels were populated by a set of linearized DNA molecules, they were imaged with automated high-resolution fluorescent microscopy using the nano*Analyzer*® (Figure 5.2B).

The size of each DNA molecule was determined by directly measuring the contour length [17]. The measured length of this clone was 50.5 micron, corresponding to 85% of the theoretical maximal stretching (complete elongation of a 183 kb DNA molecule is expected to be 59.4 microns, assuming 0.34 nm/bp). Based on measurement of 1,251 molecules, the DNA length measurement had a standard deviation of 1.3 kb (or 0.36 microns, Figure 5.2C).

By marking the positions of the fluorescent labels along the DNA molecule, the distinct distribution of the sequence motifs recognized by the nicking endonuclease was established for each fragment. Consensus sequence motif maps were constructed by comparing and clustering DNA molecules with the same sequence motif patterns. Molecules can enter into the nanochannels in either orientation; clusters that are mirror images of each other originating from the same BAC were recovered for each clone

(Figure 5.2D, top panel). A histogram representation of data from ~100 molecules is shown in the bottom panel of Figure 5.2D. The peaks represent the location of each sequence motif (GCTCTTC) along the molecules. A total of 18 nicking sites more than 1.5 kb apart were detected, and the pattern is in concordance with the reference *in silico* map.

Overall, individual molecules were labeled with 79% efficiency at nicking sites (true positives) and with a 4% false positive rate. To determine the effect of missing labels on the construction of consensus maps, various levels of coverage were tested (100 datasets of 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100x coverage) for consensus map construction. By comparing with the reference map, at 20x coverage, missing nick-labeled sites on the consensus map occurred at a rate of 0.26%. At higher coverage, there were essentially no missed label sites in the consensus. The standard deviation of consensus peak position measurements was 0.9 pixels (1 pixel corresponds to 492 bp).

Nano-mapping of 10 individual BACs was performed and the resulting sequence motif maps were highly consistent with the reference maps for the MHC region (data not shown).

### *Nano-mapping of the MHC region with 95 BACs*

We generated motif maps of the major histocompatibility (MHC) region for two haploid clone libraries from the MHC Haplotype Consortium collection. We used 49 and 46 BAC clones from the PGF and COX libraries, respectively. The DNA samples of all the clones for each library were prepared and extracted as mixtures. The two mixtures were then nick-labeled with Nt.BspQI separately and further divided into two aliquots. One aliquot from each mixture was linearized with NotI restriction enzyme and the other with BsiWI. Thus, a total of four sets of nick-labeled, linearized mixtures were loaded and imaged in the nanochannel array separately. One hundred and eight images were obtained automatically for each chip, covering 27 horizontal fields of view regions across the width (2 mm) of the array with four contiguous fields of view vertically in each region (0.5 mm) (Figure 5.3A). The contiguous images were stitched together to produce a longer field of view. Images of 23,000 molecules (3 Gb) were collected in total (sizes ranging from 20–220 kb), with a large fraction of the molecules longer than 100 kb (Figure 5.3B).

Image data from the four mixtures were combined and analyzed together, simulating a dataset obtained from a diploid DNA sample. Distances between each label were calculated for all the molecules and unsupervised clustering analysis was performed to produce a total of 140 independent clusters, each with over 100x coverage. These clusters were used to construct consensus sequence motif maps and overlapping maps were joined together to produce contig maps (Figure 5.3C). In all, we obtained 3 contigs across the 4.7 Mb MHC region (Figure 5.4). Regions harboring haplotype differences were also inspected and analyzed. The differences in nicking site distribution between the two haploid genomes were easily identified from the mixed PGF/COX dataset. All the differences identified were confirmed by analyzing the haploid PGF and COX datasets independently. As reported by Stewart et al. [25], the haploid map differences were concentrated around the HLA genes.

The maps produced by nano-mapping generally matched well with the *in silico* reference maps, but we detected discrepancies at Chr 6: 28.78-28.88 Mb and Chr 6: 30.98-31.11 Mb (Figure 5.4). At Chr 6: 28.78-28.88 Mb, there is a 4 kb insertion in the

PGF reference map relative to the COX reference map (Figure 5.5A). However, nano-mapping produced a single map that is identical to the PGF reference map, indicating that the COX reference is erroneous. This was confirmed by analyzing the PGF and COX haploid dataset separately; they had identical maps for the region. Subsequent sequencing of the clones confirmed the error in the COX reference detected by nano-mapping.

At Chr 6: 30.98-31.11 Mb, nano-mapping produced two haplotypes from the diploid dataset. One haplotype matches perfectly with the COX reference (Figure 5.5B, top panel), while the second does not completely match either the COX or PGF reference sequences (Figure 5.5B, bottom panel). Further analysis of the haploid dataset validated the top map as that derived from the COX library and that the bottom map did come from the PGF library, but with an extra nicking site within the 24 kb fragment found in the reference sequence, splitting it into two (17 kb and 7 kb). In this case, nano-mapping not only produced the two different haplotypes with ease, it also identified a nicking site left out in the original reference sequence. Sequencing of the clones reveals that an additional nick site is created by a single base difference (AAAGAGC in the reference and GAAGAGC from our sequencing data).

### *Nano-mapping for de novo sequence assembly*

To demonstrate that nano-mapping can provide useful scaffolds for *de novo* sequence assembly of short reads, we pooled the BAC clone DNA from each library and prepared two sequencing libraries for next-generation sequencing. Paired-end 100 bp sequencing of the BACs was performed, and we obtained 131 million and 142 million reads for the PGF and COX clones, respectively. For quality control, we first aligned the reads to the reference genome, hg19. Of the uniquely aligned reads, 93% of the PGF reads and 95% of the COX reads aligned to the MHC region, confirming that most BACs originated from the MHC region. The sequence motif maps produced suggested that a small subset of the BACs might be mislabeled by our supplier and did not map to the MHC region. Indeed, there was an excess of reads mapping to chromosomes 3, 7, 9, and 18, corresponding to the 5 BAC clones that were mislabeled.

*De novo* assembly of the reads was performed using SOAPdenovo [22]. Figure 5.6 shows the results of sequence assembly of a 575 kb region with the use of long-range sequence motif maps. Four sequencing contigs were oriented and placed on the *de novo* scaffold generated by nano-mapping, with a good estimate of the gap sizes between contigs (from left: 11.6 kb, 2.3 kb, 36.4 kb, and 1.2 kb, respectively). Three of the four gaps may be closed easily by designing PCR assays that bridge across them. A number of sequencing contigs could not be mapped on the sequence motif map, indicating that they were assembly errors. Blast analysis confirmed that these erroneous contigs do not properly align to the human reference sequence.

### *Nano-mapping for haplotype and structural variation analysis*

As nano-mapping produced data on molecules hundreds of thousands of base pairs in length, it is useful for long-range haplotype and structural variation analysis. Analyzing the mixed PGF and COX data, we resolved the two haplotypes across the MHC region. The most common haplotype difference detected was the presence or absence of nicking sites. An example of the presence of a nicking site in one but not the other haplotype was found at Chr6: 29.77-29.92 Mb as shown in Figure 5.7A, where the

PGF sequence had an Nt.BspQI site that was not found in the COX sequence.

Another form of variation found on these maps is that of a “shifted” nicking site. Here, a nicking site is found at one position in one haplotype but at another position in the other haplotype while the neighboring nicking sites match up perfectly (Figure 5.7B is one such example found at Chr6: 32.65-32.77 Mb). Shifting of nicking sites could be due to nearby single nucleotide variants that destroy/create nicking sites or due to single or multiple insertion/deletion events, depending on the allelic point of view.

A third type of haplotype difference identified by nano-mapping is due to insertion/deletions. Figure 5.7C shows a 5 kb insertion at Chr6: 32.41-32.53 Mb in the PGF clone that was not found in the COX clone. In this example, there were also extra nicking sites found in the PGF.

Structural variants are easily identified by nano-mapping, whether there is a haplotypic difference or not. A 30 kb tandem duplication was observed at Chr6: 31.92-31.95 Mb where the two haplotypes were identical (Figure 5.7D).

### *5.5 Discussion*

Next generation sequencing (NGS) technology has significantly advanced sequence-based genomic research and revolutionized many aspects of biological studies. However, the nature of short-read sequences makes sequence assembly difficult and unreliable, even with the introduction of paired-end sequencing. In practice, analysis of such data is often reliant on the correctness and completeness of a reference genome and is unable to resolve haplotypes or localize structural variants precisely. The importance of phase information is well-recognized [28]. Read-backed phasing is available in the Genome Analysis Tool Kit [29]; however, the resulting phased segments are short, due to the limited information provided by the short reads [30]. Misassembled contigs, misplaced and misoriented contigs, failure to join contigs, inability to accurately measure gaps in the assembled sequence, and difficulty in closing the gaps are some of the challenges in *de novo* sequence assembly of large genomes [31]. The use of sequence motif maps, especially those spanning hundreds of thousands of bases can contribute to ameliorating these difficulties. In addition to informing contig order and orientation, sequence motif maps can be used to verify contig assembly followed by disassembling incorrect contigs for reassembly in an iterative manner to improve contig fidelity.

Nano-mapping combines the specific labeling of sequence motifs (nicking endonuclease recognition sites) and automated imaging of long, uniformly stretched DNA molecules to produce sequence motif maps that are useful in resolving haplotypes, identifying the presence and location of structural variation, and providing scaffolds for *de novo* assembly of short reads from next-generation sequencers.

The fundamental advance that enables nano-mapping is the high throughput, uniform linearization of long DNA molecules compared to traditional DNA combing methods [32, 33]. The uniformity directly contributes to accurate DNA length measurement and precise distance measurements between labels. Consequently, the accuracy of the sequence motif map of each individual molecule greatly facilitates the deconvolution of mixed clones through unsupervised clustering and formation of unique and accurate consensus maps of each individual clone. Furthermore, this accuracy allows us to detect relatively small insertions/deletions, duplications and single nucleotide variants within nicking sites. The MHC haplotype maps produced here, in fact, can differentiate the two HLA-DRB1 variants (DRB*150101 and 030101) within the coding region. Differentiation of HLA-DRB1 is difficult for next generation sequencing because it is a relatively long gene with large introns in a highly repetitive region. Current approaches rely on specially designed PCR reactions [34] or target capture followed by long read sequencing [35]. However, using nano-mapping, the multiple nick sites in this gene in conjunction with adjacent sequence are sufficient to differentiate many of the HLA-DRB1 variants.

In our nick-labeling scheme, the specificity is determined by both the enzymatic nicking reaction and the fluorescent nucleotide incorporation reaction. Non-enzymatic nicking is not extended by DNA polymerase due to the lack of a functional 3'-hydroxyl group. Furthermore, the fluorescent nucleotides are covalently bound to the dsDNA without denaturation, and thus, the binding is not subjected to the variation in binding constants for some non-covalent dsDNA labeling [36-38]. The Nt.BspQI nicking endonuclease has 537 resolvable sites (having at least 1.5 kb between neighboring sites) across the 4.7 Mb MHC region in the PGF genome, yielding roughly 9 kb average spacing. At this resolution, we detected 22 haploid differences between PGF and COX in this highly variable region. However, due to the size limitation of the clones, the full phase information between the haploid map differences cannot be derived from the "diploid" dataset, even though the long-range phasing information represents a significant advance over what one can derive from next-generation sequencing alone. With additional overlap, longer genomic DNA molecules and finer resolution, a single haplotype map would be achievable across the MHC region. To construct whole genome haplotype maps, denser nick labeling (for example, by combining two nicking enzymes with two different color labels) can be employed. Alternatively, a labeling scheme, such as nick-flap labeling [26] can provide additional targeted and high-resolution sequence information besides the nicking sequence motifs for constructing a true *de novo* haplotype map.

The studies reported here demonstrate that nano-mapping can be adopted for a variety of genome analyses. Current throughput of >300 Mb per scan is sufficient for large-scale genome analysis. 20x coverage of a human genome (3.2 Gb) could be achieved in approximately 13 hours at nominal cost. Nano-mapping can be paired with next-generation sequencing to produce a fully assembled genome (after gap closing) with all classes of genetic variation characterized. Needless to say, it is also useful for analyses of genomes of other species. It can rapidly and accurately catalogue large-insert clone libraries, useful for repositories or individual labs. Nano-mapping will have special relevance in studies of new pathogens, complex metagenomics, and cancer genomes, where copy number variation and structural variation are abundant.

### *5.6 Funding sources*

This research is supported in part by NIH award to P.-Y.K. and M.X. (R01 HG005946). E.T.L was supported in part by NIH Training Grant T32 GM007175.

### *5.7 References*

1. Ley, T.J., et al., *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nature, 2008. **456**(7218): p. 66-72.
2. Siegel, A.F., et al., *Modeling the Feasibility of Whole Genome Shotgun Sequencing Using a Pairwise End Strategy.* Genomics, 2000. **68**(3): p. 237-246.

3. Fernando, M.M.A., et al., *Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis.* PLoS Genet, 2008. **4**(4): p. e1000024.
4. Sebat, J., et al., *Large-Scale Copy Number Polymorphism in the Human Genome.* Science, 2004. **305**(5683): p. 525-528.
5. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.
6. Carter, N.P., *Methods and strategies for analyzing copy number variation using DNA microarrays.* Nat Genet, 2007.
7. Medvedev, P., M. Stanciu, and M. Brudno, *Computational methods for discovering structural variation with next-generation sequencing.* Nat Meth, 2009. **6**(11s): p. S13-S20.
8. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
9. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
10. Jing, J., et al., *Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules.* Proceedings of the National Academy of Sciences, 1998. **95**(14): p. 8046-8051.
11. Zhou, S., et al., *Validation of rice genome sequence by optical mapping.* BMC Genomics, 2007. **8**(1): p. 278.
12. Zhou, S., et al., *A Single Molecule Scaffold for the Maize Genome.* PLoS Genet, 2009. **5**(11): p. e1000711.
13. Church, D.M., et al., *Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse.* PLoS Biol, 2009. **7**(5): p. e1000112.
14. Teague, B., et al., *High-resolution human genome structure by single-molecule analysis.* Proceedings of the National Academy of Sciences. **107**(24): p. 10848-10853.
15. Wu, C.-w., et al., *Optical mapping of the Mycobacterium avium subspecies paratuberculosis genome.* BMC Genomics, 2009. **10**(1): p. 25.
16. Latreille, P., et al., *Optical mapping as a routine tool for bacterial genome sequence finishing.* BMC Genomics, 2007. **8**(1): p. 321.
17. Tegenfeldt, J.O., et al., *The dynamics of genomic-length DNA molecules in 100-nm channels.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(30): p. 10979-10983.
18. Reisner, W., et al., *Statics and dynamics of single DNA molecules confined in nanochannels.* Phys Rev Lett, 2005. **94**(19): p. 196101.
19. Cao, H., et al., *Gradient nanostructures for interfacing microfluidics and nanofluidics.* Applied Physics Letters, 2002. **81**(16): p. 3058-3060.
20. Nagarajan, N., T.D. Read, and M. Pop, *Scaffolding and validation of bacterial genome assemblies using optical restriction maps.* Bioinformatics, 2008. **24**(10): p. 1229-1235.
21. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
22. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Research, 2009.

23. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**(5): p. 821-829.
24. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project.* Immunogenetics, 2008. **60**(1): p. 1-18.
25. Stewart, C.A., et al., *Complete MHC Haplotype Sequencing for Common Disease Gene Mapping.* Genome Research, 2004. **14**(6): p. 1176-1187.
26. Das, S.K., et al., *Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes.* Nucleic Acids Research, 2011. **38**(18): p. e177.
27. Xiao, M., et al., *Rapid DNA mapping by fluorescent single molecule detection.* Nucleic Acids Research, 2007. **35**(3): p. e16.
28. Tewhey, R., et al., *The importance of phase information for human genomics.* Nat Rev Genet. **12**(3): p. 215-223.
29. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet. **43**(5): p. 491-498.
30. Suk, E.-K., et al., *A comprehensively molecular haplotype-resolved genome of a European individual.* Genome Research, 2011. **21**(10): p. 1672-1685.
31. Alkan, C., S. Sajjadian, and E.E. Eichler, *Limitations of next-generation genome sequence assembly.* Nat Meth, 2011. **8**(1): p. 61-65.
32. Samad, A., et al., *Optical mapping: a novel, single-molecule approach to genomic analysis.* Genome Research, 1995. **5**(1): p. 1-4.
33. Michalet, X., et al., *Dynamic Molecular Combing: Stretching the Whole Human Genome for High-Resolution Studies.* Science, 1997. **277**(5331): p. 1518-1523.
34. Erlich, R., et al., *Next-generation sequencing for HLA typing of class I loci.* BMC Genomics. **12**(1): p. 42.
35. Pröll, J., et al., *Sequence Capture and Next Generation Resequencing of the MHC Region Highlights Potential Transplantation Determinants in HLA Identical Haematopoietic Stem Cell Transplantation.* DNA Research, 2011. **18**(4): p. 201-210.
36. Dervan, P.B. and R.W. BÃ¼rli, *Sequence-specific DNA recognition by polyamides.* Current Opinion in Chemical Biology, 1999. **3**(6): p. 688-693.
37. Felsenfeld, G. and A. Rich, *Studies on the formation of two- and three-stranded polyribonucleotides.* Biochimica et Biophysica Acta, 1957. **26**(3): p. 457-468.
38. Nielsen, P.E. and M. Egholm, *An introduction to peptide nucleic acid.* Current issues in molecular biology, 1999. **1**(1-2): p. 89-104.

# Chapter 6

# Concluding remarks and future directions

## *6.1 Summary of research findings*

The studies discussed in this dissertation involved diverse, clinically important phenotypes observed in humans and dogs. The overarching goal was to understand how DNA sequence variation contributes to these phenotypes. We made use of a variety of approaches according to the specific hypotheses associated with the phenotypes. The approaches we took and complementary and competing approaches were briefly discussed in Chapter 1, which also serves to provide a historical background of how new technologies have driven progress in the field of genetics/genomics.

In Chapter 2, I described the use of Affymetrix's MitoChip platform to sequence the mitochondrial genomes of hundreds of individuals. The focused design of the MitoChip was well suited for us to explore the hypothesis that mitochondrial variants contribute to cancer risk and ageing phenotypes. Our pancreatic cancer study represents the largest comprehensive mitochondrial sequencing study with a case-control design to date. Despite the respectable sample size, it was evident that more rare or singleton variants remain to be found, many of which likely impact mitochondrial function. Because of the importance of the proteins encoded by the mitochondrial genome in oxidative phosphorylation, mitochondrial variants might be associated with many other phenotypes not discussed here as well. Sequencing reads from the mitochondrial genome are often discarded during the analysis of next-generation sequencing data. We may benefit from mining existing exome and whole-genome datasets and be able to identify additional functionally relevant variants. In particular, next-generation sequencing data may help us examine heteroplasmy levels and resolve clonal structures of mitochondria that are obscured in array data.

In Chapter 3, I described our study using emulsion PCR and next-generation sequencing for targeted sequencing of 95 genes involved in chemotherapy drug pathways. We took advantage of the family structure in the samples to identify variants that could contribute to chemotherapy response. Our PCR-based candidate gene approach could be applied in the clinical setting as PCR provides specific amplification of regions of interest and next-generation sequencing enables high-throughput sequencing of many loci simultaneously. It is important to note that we observed large variation in drug response in non-tumor cells, and the use of lymphoblastoid cell lines allowed us to survey natural variation that would contribute to drug response independent of cancer-related changes.

I described our work on fine-mapping a region associated with adult-onset deafness in Border collies in Chapter 4. Despite having a strong, definitive signal in the initial genome-wide association study, which variant was causing the phenotype was unclear. Therefore, we relied on direct sequencing of the candidate region for hypothesis generation. By using solution-phase target capture and next-generation sequencing, we were able to sequence the region to high coverage in a small sample of carefully selected dogs. I believe we have identified strong candidates that warrant functional follow-up studies.

The targeted sequencing approaches mentioned usually require special setups. For example, one needs the GeneChip module in order to scan the Affymetrix microarrays.

Also, in order to use RainDance Technologies' emulsion PCR platform, access to either RDT1000 or the newer ThunderStorm system is required, thus posing a challenge and a barrier to entry for those without access. In contrast, solution-phase target capture requires no more than typical labware. This should be taken into consideration when deciding which approach would be the most appropriate.

Chapter 5 differs from the other chapters in that the goal of the study was to develop an optical mapping technique so that it could be used in structural variation analysis and *de novo* sequence assembly. Although important, these two areas of research were not discussed in depth in the previous chapters. Thus far, we have focused on single-base changes; however, it is conceivable as well as likely that structural variation impacts our phenotypes of interest. Structural variants have been discovered through the use of microarrays and next-generation sequencing, but it is likely that some are missed due to the limitations of these techniques. Optical mapping potentially offers another layer of information that could benefit structural variation analysis. At the same time, optical mapping data could provide physical mapping data with a higher resolution than traditional physical mapping techniques. To avoid having to rely on the presence of a cloned BAC library, we will adopt our method to genomic DNA. We anticipate additional bioinformatics challenges, but we believe that sequence assembly would be made easier with shotgun optical mapping data generated from genomic DNA.

### *6.2 The future of high-throughput genomics*

It is abundantly clear that we can detect single-base change and simple structural variation in a high-throughput manner using microarrays and next-generation sequencing.

Future development would likely be shifted towards improving our ability to decipher long-range haplotype patterns and analyze the data in a more reference-independent fashion. The current trend suggests that sequencing reads will only be at best in the kilobase range in the near future. This means that information about the relationship between two distant alleles on the same chromosome is often lost. There are bioinformatic means to infer haplotypes, but the process is error-prone. Because DNA molecules of hundreds of kilobases are used in optical mapping, long-range information is preserved. This information would benefit both haplotype analysis and sequence assembly when neighboring contigs have to be joined.

As the amount of data continues to grow exponentially, bioinformatic support becomes more and more important. For large datasets, manual analysis on an Excel spreadsheet is no longer preferred or even possible. Data storage also becomes an issue, and there have been attempts to compress the data being generated without or with minimal information loss. New bioinformatics tools are needed so that researchers with less bioinformatics training can access the large amount of data that is openly available. There are several popular next-generation sequencing platforms that provide complement data; therefore, standardization in sequence and variant data format is essential. BAM files and VCF files are now standard for alignment files and variant files, respectively. However, standardization for cataloging structural variants is complicated by the fact that an insertion could be described as a deletion depending on the allele considered. Also, there are complex structural variants that involve multiple events. Having a well-documented system for all types of variants will be integral to downstream analysis.

Nonetheless, data are being generated faster than can be analyzed. In a routine exome sequencing experiment, one is likely to find millions of variants with alleles different from the reference. Inevitably, a portion of those would be predicted to have a functional impact. Traditional cell-based mutagenesis studies are too labor-intensive for validation of found variants. Being able to functionally validate these variants would require development of high-throughput assays that could accommodate the large number of "interesting" variants.

This dissertation has focused on the use of genomic tools to study variation on the DNA level. However, it should be noted that these technologies have also enabled for example epigenetic and gene expression studies. Most recently, next-generation sequencing has been extensively used to study the transcriptomes and methylation patterns of different organisms. This is important because DNA sequence variation represents only one source of variation that could contribute to a phenotype.

Ultimately, one is left to wonder if advances in the basic research setting will be translated to the clinical setting. Wide adoption of these new technologies in the clinic remains to be seen, but next-generation sequencing is slowly making its way into diagnostic labs. Replacing Sanger sequencing, next-generation sequencing will shorten the sequencing turnaround time and enable of sequencing of many loci in a single assay. This will likely speed up diagnosis and benefit the patient. As the cost of sequencing continues to drop, one might wonder whether the promises of personalized medicine will finally be realized in the near future.

**Table 2.1 Characteristics of pancreatic cancer cases and controls in a population based study in the San Francisco Bay Area, California (1995-1999).**

| | | Cases | Controls |
|---|---|---|---|
| | | N = 286 | N = 283 |
| Age (years) | Mean (s.d.) | 65 (11) | 64 (12) |
| | | n (%) | n (%) |
| Education (years) | | | |
| | 1-12 | 120 (42) | 91 (32) |
| | >12-16 | 111 (39) | 120 (42) |
| | >16 | 55 (19) | 72 (25) |
| Self-reported race/ethnicity | | | |
| | White, non-hispanic | 226 (79) | 223 (79) |
| | White, hispanic | 16 (6) | 20 (7) |
| | Black | 22 (8) | 11 (4) |
| | Asian | 17 (6) | 22 (8) |
| | Other | 5 (2) | 7 (2) |
| Sex | | | |
| | Men | 153 (54) | 152 (54) |
| | Women | 133 (46) | 131 (46) |
| Cigarette smoking | | | |
| | Never smoked | 82 (29) | 106 (39) |
| | Former smoker | 131 (48) | 129 (48) |
| | Current smoker | 64 (23) | 35 (13) |
| Halogroup N | | 246 (87) | 248 (88) |
| Haplogroup L | | 27 (10) | 14 (5) |
| Haplogroup M | | 11 (4) | 21 (7) |

**Table 2.2 Odds Ratios (OR) and 95% confidence intervals (CI) for pancreatic cancer associated with haplogroup N subgroups and common mtDNA variants among haplogroup N participants, San Francisco Bay Area, California (1995-1999).**

| | | Cases | Controls | | |
|---|---|---|---|---|---|
| Haplogroup N subgroups | | N = 246 | N = 248 | OR (95 % CI)[a] | P-value[c] |
| | | n (%) | n (%) | | |
| | H | 107 (44) | 101 (40) | 1.0 (Ref.) | |
| | V | 8 (4) | 8 (3) | 0.92 (0.3 - 2.8) | 0.88 |
| | J | 27 (11) | 24 (9) | 1.03 (0.52 - 2.0) | 0.94 |
| | T | 24 (10) | 27 (11) | 0.6 (0.31 - 1.2) | 0.13 |
| | U | 31 (13) | 35 (14) | 0.67 (0.37 - 1.2) | 0.17 |
| | K | 10 (4) | 24 (9) | 0.32 (0.13 - 0.76) | 0.01 |
| | B, F | 8 (3) | 9 (3) | 0.86 (0.3 - 2.42) | 0.77 |
| | A, I, W, X, Y | 22 (8) | 27 (9) | 0.89 (0.45 - 1.8) | 0.73 |
| | | | | | |
| Region/gene[b] | Site, allele | | | | |
| Complex I | | | | | |
| ND2 | mt5460g | 20 (8) | 6 (2) | 3.9 (1.5 - 10) | 0.004 |
| Complex IV | | | | | |
| COIII | mt9698c | 10 (4) | 24 (9) | 0.44 (0.20 - 0.97) | 0.02 |
| 16S | mt1811g | 20 (7) | 40 (14) | 0.51 (0.29 - 0.91) | 0.008 |
| tRNA | mt12307g | 16 (7) | 34 (14) | 0.45 (0.24 - 0.85) | 0.02 |
| HV2 | mt150t | 13 (5) | 26 (11) | 0.70 (0.49 - 0.99) | 0.03 |

[a] All analyses were adjusted for age, sex and 6 eigenvectors of mitochondrial genetic ancestry derived from principal component analysis.

[b] Nominally significant results for 5 common variants out of 66 total common variants detected.

[c] Multiple comparisons adjusted p-value (a=0.05) for 8 haplogroups (a=0.006) and 66 common variants (a=0.0008).

**Table 2.3 Rare, haplogroup specific mtDNA variants associated with pancreatic cancer, San Francisco Bay Area, California (1995-1999).**

| Region | Gene | Site | Cases | Controls | Haplogroup | Common allele | Rare allele | Amino Acid position | PhastCons | PhyloP | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Complex I | ND4L | 10586 | 7 | 1 | L | tcg | tca | S39S | 0.00 | -7.02 | 0.046 |
| | ND5 | 12810 | 6 | 0 | L | tga | tgg | W158W | 0.00 | -0.58 | 0.046 |
| | ND5 | 12954 | 4 | 0 | N | gct | gcc | A206A | 0.00 | -2.44 | 0.040 |
| | ND5 | 13485 | 6 | 0 | L | ata | atg | M383M | 0.00 | -4.18 | 0.046 |
| | **ND5** | **14000** | **6** | **0** | **L** | **cta** | **caa** | **L555Q**[a] | 0.00 | -0.50 | **0.046** |
| Complex III | **CytB** | **14763** | **0** | **3** | **L** | **aaa** | **aca** | **K6N**[a] | 1.00 | 4.47 | **0.003** |
| | CytB | 14869 | 0 | 3 | L | ctg | cta | L41L | 0.00 | -0.59 | 0.01 |
| | CytB | 15784 | 7 | 1 | N | cct | ccc | P346P | 0.00 | -3.64 | 0.040 |
| Complex IV | COI | 5951 | 6 | 0 | L | gga | ggg | G16G | 0.32 | -1.60 | 0.046 |
| | COI | 6071 | 6 | 0 | L | gtt | gtc | V56V | 0.00 | -8.27 | 0.046 |
| | COI | 6260 | 1 | 6 | N | gag | gaa | E119E | 0.98 | -0.11 | 0.050 |
| | COIII | 9548 | 8 | 2 | N | ggg | gga | G114G | 0.00 | -6.52 | 0.050 |
| Complex V | ATP6 | 9072 | 6 | 0 | L | tca | tcg | S182S | 0.00 | -1.59 | 0.046 |
| rRNA | 12S | 951 | 1 | 6 | N | g | a | | 0.00 | -2.22 | 0.050 |
| | 12S | 961 | 3 | 0 | N | t | g | | 0.00 | -4.81 | 0.050 |
| HV2 | | 193 | 6 | 0 | N | c | t | | 0.73 | 0.18 | 0.040 |
| | | 296 | 8 | 0 | L | a | g | | 0.01 | -0.12 | 0.046 |
| | | 316 | 6 | 0 | L | g | a | | 0.03 | -1.32 | 0.046 |
| Noncoding | | 16527 | 10 | 1 | N | c | t | | 0.00 | -1.40 | 0.020 |

[a] Nonsynonomous variant predicted to have a damaging effect on the resulting protein using PolyPhen2.

**Table 2.4 Detailed analysis of nonsynonymous substitutions unique to sedentary (AEE < 401 kcal/d) and active (AEE > 907 kcal/d) Health ABC Study participants.** Significant prediction (SIFT, PolyPhen) and conservation (PhastCons, PhyloP) scores are indicated in bold.

| Complex | I | | | | | | | | | | | | | | | | III | | | | | | IV | | V | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | ND1 | | | | ND2 | | | ND3 | ND4 | | ND5 | | | | | ND6 | CytB | | | | | | COI | COIII | ATP8 | |
| Nt. position | 3308t | 3593t | 3943a | 4172t | 4501c | 4640a | 4890a | 10345t | 11065a | 11172a | 12634a | 12811t | 12952g | 13676a | 13934c | 14180t | 14927a | 15257g | 15317g | 15326g | 15758a | 15866a | 7080t | 9300g | 8393c | 8490t |
| AA Position | M1 | V96 | I213 | L289 | S11 | I57 | I141 | I96 | L102 | N138 | I100 | Y159 | A206 | N447 | T533 | Y165 | T61 | D171 | A191 | T194 | I338 | N374 | F393 | A32 | P10 | M42 |
| AEE kcal/day | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 230 | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 291 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . |
| 307 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | A | . | . | . | . | . | . |
| 316 | . | . | . | . | . | . | . | . | . | . | . | H | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 332 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 341 | . | . | . | . | . | . | . | . | R | . | . | . | . | . | . | . | . | . | . | . | V | . | . | . | . | . |
| 351 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | N | . | . | . | . | . | . | . | . |
| 366 | . | . | . | Q | . | . | . | . | . | . | . | . | . | . | . | . | A | . | . | . | . | . | . | . | . | . |
| 381 | . | . | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 391 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | T | . | . | . | . | . | . | . |
| 396 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | V | . | . | . | . | . |
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 935 | . | A | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 956 | . | . | . | . | . | . | . | . | . | . | . | . | T | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 985 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | D | . | . | . | . |
| 1010 | . | . | . | . | . | . | V | . | . | . | . | . | . | S | . | . | . | . | . | . | . | . | L | . | . | . |
| 1042 | . | . | . | . | . | M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | T | . | T |
| 1042 | R | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 1048 | . | . | . | . | . | . | . | T | . | . | . | . | . | . | . | C | . | . | . | . | . | . | . | . | . | . |
| 1061 | . | . | . | . | F | . | . | . | . | . | Y | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 1081 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | M | . | . | . | . | . | . | . | . | . | . | . |
| 1141 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | M | . | . | . | . | . | . | . | . | . | . | . |
| PhastCons | **0.69** | **0.00** | **0.97** | **0.12** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.23** | **0.95** | 0.00 | 0.00 | **0.11** | 0.00 | **0.69** | **0.16** | **0.99** | **0.67** | 0.00 | 1.00 | 1.00 | **0.89** | 0.00 | **0.03** | 0.00 |
| PhyloP | **2.23** | **1.17** | **4.00** | **2.05** | -1.19 | -3.05 | -0.07 | -1.16 | -8.02 | **2.08** | **4.39** | -0.08 | -0.20 | **1.23** | -0.27 | **1.69** | **0.77** | **2.74** | **3.85** | -1.08 | 1.81 | 1.91 | **2.43** | -0.29 | -0.09 | -0.72 |
| SIFT | **0.00** | **0.03** | 1.00 | **0.00** | 0.17 | **0.00** | 1.00 | 0.29 | **0.00** | **0.08** | **0.00** | 0.71 | 0.62 | **0.05** | 1.00 | **0.03** | **0.02** | **0.04** | 0.11 | 0.83 | **0.00** | **0.00** | 0.25 | **0.00** | **0.10** | 0.75 |
| PolyPhen | 0.13 | 0.04 | 0.00 | **0.99** | 0.00 | **0.55** | **0.82** | 0.00 | **1.00** | **0.98** | **0.98** | 0.00 | 0.00 | 0.09 | 0.00 | **0.91** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.83** | 0.00 | 0.00 | **0.90** | 0.07 |

**Table 2.5 Rare variant burden tests of associations across hypervariable region 2 for metabolic rate and energy expenditure in the Health ABC Study.**

| | N | $P_{T1}$ | $P_{T5}$ | $P_{WE}$ | $P_{VT}$ |
|---|---|---|---|---|---|
| RMR[1] | 135 | 0.64 | 0.86 | 0.86 | 0.90 |
| TEE[2] | 129 | 0.20 | 0.14 | 0.20 | 0.32 |
| AEE[3] | 129 | 0.09 | 0.01 | 0.02 | 0.03 |
| PAL[4] | 129 | 0.13 | 0.006 | 0.01 | 0.02 |

[1] Resting metabolic rate (RMR) was measured via indirect calorimetry.
[2] Total energy expenditure (TEE) was measured using the 2-point doubly-labeled water technique.
[3] Activity energy expenditure (AEE) was calculated as [(TEE*0.90) − RMR].
[4] Physical activity level (PAL) was calculated as TEE/RMR.
P-values for T1 (1% allele-frequency threshold), T5 (5% allele-frequency threshold), WE (weighted), and VT (variable threshold), analyses are displayed. A significance level of $p \leq 0.01$ is used after multiple testing correction (a=0.05) for 9 mtDNA regions, based on 10,000 independent simulations.

**Table 2.6 Characteristics of dementia cases and controls among sequenced Health ABC participants (n=135).**

| | *No dementia* | *Dementia* |
|---|---|---|
| N (%) | 113 (84) | 22 (16) |
| Age, mean (SD) | 73.2 (2.8) | 74.7 (3.0) |
| BMI (kg/m$^2$) , mean (SD) | 26.8 (4.6) | 25.8 (5.8) |
| APOEe4 carrier, n (%) | 28 (25) | 7 (32) |
| Prevalent diabetes, n (%) | 21 (19) | 5 (23) |
| Sex, n (%) | | |
| Male | 55 (48) | 8 (38) |
| Female | 59 (52) | 13 (62) |
| Education, n (%) | | |
| Less than HS | 15 (13) | 4 (19) |
| HS grad | 35 (31) | 9 (43) |
| Postsecondary | 63 (56) | 8 (38) |
| Haplogroup, n* (%) | | |
| H | 51 (45) | 9 (41) |
| U | 9 (8) | 2 (9) |
| K | 13 (11) | 3 (14) |
| T | 14 (12) | 4 (18) |
| J | 10 (9) | 3 (14) |
| V | 2 (2) | - |
| I | 5 (4) | - |
| W | 1 (1) | - |
| X | 3 (3) | - |

*Numbers do not add up to total due to missing information for haplogroups.

**Table 2.7 Nonsynonymous substitutions, tRNA, rRNA, and HV2 region variants unique to Health ABC Study participants with dementia.**
Values in bold indicate sites that are predicted to significantly impact *in-silico* modeling of evolutionary conservation (PhastCons and PhyloP >0), protein stability (SIFT <0.1) and function (PolyPhen 'damaging' prediction).

| | | Complex | I | | | | | | III | | IV | V | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Function | ND1 | ND2 | | ND4L | ND5 | | CytB | | COI | ATP8 | tRNA | | | HV2 | | 16S | |
| | | Nucleotide | m.3943, A>G | m.4890, A>G | m.5461, C>A | m.10750, A>G | m.12811, T>C | m.13676, A>G | m.15317, G>A | m.15326, A>G | m.7080, T>C | m.8519, G>A | m.5527, A>G | m.5567, T>C | m.5592, A>G | m.114, C>T | m.238, A>T | m.1700, T>C | m.2141, T>C |
| | | Protein | p.I213V | p.I141V | p.A331D | p.N94S | p.Y159H | p.N447S | p.A191T | p.T194A | p.F393L | p.E52K | | | | | | | |
| Sex, Age | Haplogroup | | | | | | | | | | | | | | | | | | |
| F, 72y | H | | . | . | . | . | H | . | . | . | . | . | . | . | . | . | . | . | . |
| M, 69y | K | | . | . | . | . | . | . | . | . | . | . | . | . | g | . | . | . | . |
| F, 71Y | U | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | c | . |
| M, 77y | H | | . | V | . | . | . | S | . | . | L | . | . | c | . | . | . | . | . |
| F, 74y | H | | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| F, 77y | T | | . | . | . | S | . | . | . | . | . | . | . | . | . | . | . | . | . |
| F, 73y | H | | . | . | . | . | . | . | . | A | . | . | . | . | . | . | . | . | . |
| F, 74y | I | | . | . | . | . | . | . | . | . | . | K | . | . | . | . | . | . | . |
| F, 79y | J | | . | . | D | . | . | . | . | . | . | . | . | . | . | . | t | . | . |
| M, 76y | K | | . | . | . | . | . | . | . | . | . | . | . | . | . | t | . | . | . |
| M, 73y | T | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | c |
| F, 75y | T | | . | . | . | . | . | . | . | . | . | . | g | . | . | . | . | . | . |
| F, 72y | J | | . | . | . | . | . | . | T | . | . | . | . | . | . | . | . | . | . |
| | | PhastCons | **0.97** | 0.00 | 0.00 | **1.45** | 0.00 | **0.11** | **0.67** | 0.00 | **0.89** | **0.41** | 0.00 | 0.00 | **0.98** | 0.01 | **0.43** | 0.00 | 0.00 |
| | | PhyloP | **4.00** | -0.07 | **0.20** | **0.98** | -0.08 | **1.23** | **3.85** | -1.08 | **2.43** | **0.54** | -1.98 | -0.18 | **0.07** | -2.32 | **0.87** | -3.41 | -2.61 |
| | | SIFT | 1.00 | 1.00 | 0.33 | 0.25 | 0.71 | **0.05** | 0.11 | 0.83 | 0.25 | **0.09** | | | | | | | |
| | | PolyPhen | 0.00 | **0.82** | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 0.00 | 0.49 | | | | | | | |

**Table 3.1 Genes being resequenced and drug pathways for which they are relevant.** (5FU: 5-flourouracil; Pt.: platinum; Tax.: taxanes; Campt.: camptothecin; Dox.: doxorubicin)

| ***Gene*** | ***5FU*** | ***Pt.*** | ***Tax.*** | ***Campt.*** | ***Dox.*** |
|---|---|---|---|---|---|
| ABCB1 | | | + | + | + |
| ABCC1 | | | + | + | + |
| ABCC2 | | + | + | + | + |
| ABCC4 | | | | + | |
| ABCG2 | | + | + | + | + |
| ADPRT | | | | + | |
| ATP7A | | + | | | |
| ATP7B | | + | | | |
| BCHE | | | | + | |
| CDC45L | | | | + | |
| CES1 | | | | + | |
| CES2 | | | | + | |
| CFLAR | + | | | | |
| CSAG2 | | | + | | |
| CYP1B1 | | | + | | |
| CYP2C8 | | | + | | |
| CYP3A4 | | | + | + | + |
| CYP3A5 | | | + | + | + |
| DHFR | + | | | | |
| DPYD | + | | | | |
| DPYS | + | | | | |
| DTYMK | + | | | | |
| DUT | + | | | | |
| ECGF1 | + | | | | |
| ERCC1 | | + | | | |
| ERCC2 | | + | | | + |
| ERCC3 | | + | | | |
| ERCC4 | | + | | | |
| ERCC6 | | + | | | |
| FASLG | + | | | | |
| FDXR | + | | | | |
| FPGS | + | | | | |
| GGH | + | | | | |
| GSTM1 | | + | | | |
| GSTP1 | | + | | | + |
| GSTT1 | | + | | | |
| HMGB1 | | + | | | |
| MAP4 | | | + | | |

| | | | | | |
|---|---|---|---|---|---|
| MAPT | | | + | | |
| MLH1 | | + | | | + |
| MPO | | + | | | |
| MSH2 | | + | | | + |
| MT1A | | + | | | |
| MT2A | | + | | | |
| MTHFR | + | | | | |
| NFKB1 | + | | | + | |
| NME1 | + | | | | |
| NME2 | + | | | | |
| NQO1 | | + | | | + |
| NR1I2 | | | + | | |
| NT5C | + | | | | |
| PMS2 | | + | | | |
| PNKP | | | | + | |
| POLB | | + | | | |
| POLH | | + | | | |
| POLM | | + | | | |
| REV3L | | + | | | |
| RRM1 | + | | | | |
| RRM2 | + | | | | |
| SLC31A1 | | + | | | |
| SLCO1B1 | | | | + | |
| SLCO6A1 | | + | | | |
| SMARCA1 | | + | | | |
| SOD1 | | + | | | |
| TDP1 | | | | + | |
| TK1 | + | | | | |
| TOP1 | | | | + | + |
| TP53 | + | | | | |
| TUBA1B | | | + | | |
| TUBA2 | | | + | | |
| TUBB | | | + | | |
| TYMS | + | | | | |
| UCK2 | + | | | | |
| UGT1A1 | | | | + | |
| UGT1A3 | | | | + | |
| UGT1A4 | | | | + | |
| UGT1A7 | | | | + | |
| UGT1A8 | | | | + | |
| UGT1A9 | | | | + | |
| UGT2B15 | | | | + | |
| UGT2B7 | | | | + | |

| | | | | | |
|---|---|---|---|---|---|
| UMPS | + | | | | |
| UNG | + | | | | |
| UPB1 | + | | | | |
| UPP1 | + | | | | |
| XPA | | + | | | |
| XRCC1 | | | | + | |

**Table 3.2 Gene annotations based on molecular function and biological processes.** The annotations were based on Panther (Protein ANalysis THrough Evolutionary Relationship) annotations.

| *Gene* | *Molecular function* | *Biological process(es) involved* |
|---|---|---|
| ABCB1 | ABC transporter | Extracellular transport and import |
| ABCC1 | ABC transporter | Small molecule transport; Detoxification |
| ABCC2 | ABC transporter | Small molecule transport; Detoxification |
| ABCC4 | ABC transporter | Small molecule transport; Detoxification |
| ABCG2 | Transporter | Transport |
| ATP7A | Cation transporter | Cation transport; Calcium ion homeostasis |
| ATP7B | Cation transporter | Cation transport; Calcium ion homeostasis |
| BCHE | Esterase | Neuromuscular synaptic transmission |
| CDC45L | Replication origin binding protein | DNA replication; DNA replication |
| CES1 | Esterase | Detoxification |
| CES2 | Esterase | Detoxification |
| CFLAR | Cysteine protease | Proteolysis; Apoptosis |
| CSAG2 | Unclassified | Unclassified |
| CYP1B1 | Oxygenase | Fatty acid and steroid metabolism |
| CYP2C8 | Oxygenase | Fatty acid and steroid metabolism |
| CYP3A4 | Oxygenase | Steroid hormone metabolism |
| CYP3A5 | Oxygenase | Steroid hormone metabolism |
| DHFR | Unclassified | Pyrimidine metabolism; DNA metabolism |
| DPYD | Dehydrogenase | Pyrimidine metabolism |
| DPYS | Hydrolase | Nucleic acid metabolism |
| DTYMK | Nucleotide kinase | DNA metabolism |
| DUT | Phosphatase; Hydrolase | Nucleic acid metabolism |
| ECGF1 | Glycosyltransferase | Pyrimidine metabolism |
| ERCC1 | Endodeoxyribonuclease | DNA repair |
| ERCC2 | DNA helicase | DNA repair |
| ERCC3 | DNA helicase; Hydrolase | DNA repair |
| ERCC4 | Endodeoxyribonuclease | DNA repair |
| ERCC6 | DNA helicase | DNA repair |
| FASLG | Cytokine | Ligand-mediated signaling; Apoptosis |
| FDXR | Reductase | Ferredoxin metabolism |
| FPGS | Synthase; Ligase | Carbon metabolism |
| GGH | Cysteine protease | Vitamin metabolism |
| GPX1 | Peroxidase | Detoxification; Free radical removal |
| GSTM1 | Transferase | Detoxification |
| GSTP1 | Transferase | Detoxification |
| GSTT1 | Transferase; Epimerase/racemase | Detoxification; Free radical removal |

| | | |
|---|---|---|
| HMGB1 | Unclassified | Unclassified |
| MAP4 | Microtubule binding protein | Apoptosis; Cell structure |
| MAPT | Microtubule binding protein | Apoptosis; Cell structure |
| MLH1 | DNA-binding protein | DNA repair; Meiosis; Oncogenesis |
| MPO | Peroxidase | Granulocyte-mediated immunity |
| MSH2 | Damaged DNA-binding protein | DNA repair; Meiosis |
| MT1A | Unclassified | Unclassified |
| MT2A | Unclassified | Unclassified |
| MTHFR | Reductase | Coenzyme metabolism |
| NFKB1 | Transcription factor | Immunity |
| NME1 | Nucleotide kinase | Pyrimidine metabolism |
| NME2 | Nucleotide kinase | Pyrimidine metabolism |
| NQO1 | Unclassified | Unclassified |
| NR1I2 | Transcription factor | Steroid hormone-mediated signaling |
| NT5C | Unclassified | Unclassified |
| PARP1 | Glycosyltransferase | DNA repair; Protein ADP-ribosylation |
| PMS2 | DNA-binding protein | DNA repair |
| PNKP | Nucleotide kinase/phosphatase | DNA repair |
| POLB | DNA-directed DNA polymerase | DNA repair |
| POLH | DNA-directed DNA polymerase | DNA repair |
| POLM | DNA-directed DNA polymerase | DNA recombination; Immunity |
| REV3L | DNA-directed DNA polymerase | DNA repair |
| RRM1 | Reductase | Purine metabolism; Pyrimidine metabolism |
| RRM2 | Unclassified | Unclassified |
| SLC31A1 | Cation transporter | Cation transport |
| SLCO1B1 | Transporter | Anion transport |
| SLCO6A1 | Transporter | Anion transport |
| SMARCA1 | DNA helicase | mRNA transcription regulation |
| SOD1 | Oxidoreductase | Immunity and defense |
| TDP1 | Phosphodiesterase | DNA repair |
| TK1 | Nucleotide kinase | Pyrimidine metabolism |
| TOP1 | DNA topoisomerase | DNA replication |
| TP53 | Transcription factor | DNA repair; Induction of apoptosis |
| TUBA1B | Tubulin | Chromosome segregation |
| TUBA2 | Tubulin | Chromosome segregation |
| TUBB | Tubulin | Chromosome segregation |
| TYMS | Synthase; Methyltransferase | Pyrimidine metabolism |
| UCK2 | Nucleotide kinase | Pyrimidine metabolism |
| UGT1A1 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |

| | | |
|---|---|---|
| UGT1A3 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT1A4 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT1A7 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT1A8 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT1A9 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT2B15 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UGT2B7 | Glycosyltransferase | Polysaccharide/steroid hormone metabolism |
| UMPS | Unclassified | Unclassified |
| UNG | DNA glycosylase; Hydrolase | Carbohydrate metabolism; DNA repair |
| UPB1 | Hydrolase | Carbon metabolism |
| UPP1 | Phosphorylase | Pyrimidine metabolism |
| XPA | Damaged DNA-binding protein | DNA repair |
| XRCC1 | Nucleic acid binding | DNA repair |

**Table 3.3 Sources of data.**

| | Reads | Gigabases | %Data |
|---|---|---|---|
| ILMN (35-nt Reads) | 348 M | 12.2 | 15% |
| Davis (40-nt Reads) | 29 M | 1.2 | 1% |
| GCF (36-nt Reads) | 1,958 M | 70.5 | 84% |
| Total | 2,335 M | 83.8 | |

**Table 4.1 Sample summary.**
Samples for the primary genome-wide association study (GWAS) and targeted genotyping were collected from two countries, with breakdown of cases and controls provided for a total of 405 Border collies.

| Sample | Country of Origin | Cases | Controls | Total |
|---|---|---|---|---|
| Primary GWAS | USA | 20 | 28 | 48 |
| Follow-up genotyping | USA | 0 | 14 | 14 |
| | Finland | 3 | 59 | 62 |
| Replication | USA | 16 | 265 | 281 |
| **Totals** | | 39 | 366 | 405 |

**Table 4.2 List of predicted genes targeted for target capture sequencing and probe coverage by gene.** ("n/a" is used where a gene is within another predicted gene.)

| Gene ID | Gene name | Targeted region | No. of target regions within gene |
|---|---|---|---|
| NM_006040 | *HS3ST4* | chr6:23237568-23638888 | 24 |
| NM_001012981 | *ZKSCAN2* | chr6:23976533-23997352 | 25 |
| NM_001169 | *AQP8* | chr6:24003232-24012663 | 9 |
| NM_016309 | *LCMT1* | chr6:24046309-24086693 | 48 |
| NM_001076019 | *C8H9orf30* | chr6:24057127-24060285 | n/a |
| NM_001016739 | *c9orf30* | chr6:24057154-24059998 | n/a |
| NM_018054 | *ARHGAP17* | chr6:24099958-24196585 | 101 |
| NM_052944 | *SLC5A11* | chr6:24201588-24237597 | 45 |
| NM_014494 | *TNRC6A* | chr6:24265253-24373846 | 114 |
| NM_006910 | *RBBP6* | chr6:24498008-24531951 | 30 |
| NM_006539 | *CACNG3* | chr6:24644639-24732130 | 97 |
| NM_002738 | *PRKCB* | chr6:24760568-25086234 | 389 |
| NM_022097 | *CHP2* | chr6:25148807-25153168 | 4 |
| NM_033266 | *ERN2* | chr6:25174871-25190051 | 12 |
| NM_005030 | *PLK1* | chr6:25188004-25201121 | 11 |
| NM_032486 | *DCTN5* | chr6:25208311-25234185 | 26 |
| NM_024675 | *PALB2* | chr6:25232485-25262468 | 31 |
| NM_005003 | *NDUFAB1* | chr6:25265061-25276929 | 18 |
| NM_019116 | *UBFD1* | chr6:25282922-25300748 | 12 |
| NR_003501 | *EARS2* | chr6:25298942-25322502 | 30 |
| NM_015044 | *GGA2* | chr6:25340027-25369230 | 24 |
| NM_153603 | *COG7* | chr6:25378018-25460052 | 73 |
| NM_000336 | *SCNN1B* | chr6:25462920-25487946 | 20 |
| NM_001039 | *SCNN1G* | chr6:25593643-25618204 | 27 |
| NM_020718 | *USP31* | chr6:25649946-25722762 | 78 |
| NM_006043 | *HS3ST2* | chr6:25842382-25935014 | 107 |
| NM_144672 | *OTOA* | chr6:26134411-26197802 | 78 |
| NM_001077180 | *METTL9* | chr6:26208563-26245138 | 33 |
| NM_030691 | *Igsf6* | chr6:26216896-26223970 | n/a |
| NM_001802 | *CDR2* | chr6:26324472-26345913 | 19 |
| NM_018119 | *POLR3E* | chr6:26356857-26387924 | 29 |
| NM_013302 | *EEF2K* | chr6:26396174-26463173 | 69 |
| NM_173615 | *VWA3A* | chr6:26500885-26562226 | 58 |
| NM_001164579 | *C16orf52* | chr6:26568920-26637167 | 71 |
| NM_173806 | *PDZD9* | chr6:26642205-26656344 | 13 |
| NM_003366 | *UQCRC2* | chr6:26655187-26682978 | 24 |
| NM_026458 | *Abca14* | chr6:26703127-26835863 | 63 |
| NR_024051 | *Abca16* | chr6:26887834-27195268 | 155 |
| NM_001888 | *CRYM* | chr6:27213760-27232848 | 21 |
| NM_145865 | *ANKS4B* | chr6:27236043-27246352 | 9 |

|  |  |  |  |
|---|---|---|---|
| NM_003460 | *ZP2* | chr6:27256838-27269770 | 13 |
| NM_020422 | *TMEM159* | chr6:27282092-27294013 | 13 |
| NM_017539 | *DNAH3* | chr6:27304543-27466095 | 157 |
| NM_020424 | *LYRM1* | chr6:27467260-27489430 | 20 |
| NM_173475 | *DCUN1D3* | chr6:27487425-27525621 | 32 |
| NM_030941 | *LOC81691* | chr6:27532246-27578425 | 35 |
| NM_001142725 | *ERI2* | chr6:27576410-27587378 | 8 |
| NM_005622 | *ACSM3* | chr6:27584660-27605071 | 20 |
| NM_017736 | *THUMPD1* | chr6:27633650-27644499 | 8 |
| NM_178414 | *Acsm4* | chr6:27643513-27671891 | 33 |
| NM_001087266 | *nono* | chr6:27683924-27687135 | 1 |
| NR_003277 | *LOC728643* | chr6:27687755-27690305 | 2 |
| NM_001101952 | *ACSM2A* | chr6:27892180-27921107 | 22 |
| NM_017888 | *ACSM5* | chr6:27934924-27967023 | 24 |
| NM_174924 | *PDILT* | chr6:27970602-28014520 | 45 |
| NM_003361 | *UMOD* | chr6:28018265-28035478 | 22 |
| NM_001502 | *GP2* | chr6:28038975-28055408 | 21 |
| NM_001002911 | *GPR139* | chr6:28276437-28318406 | 48 |
| NM_016235 | *GPRC5B* | chr6:28439878-28462832 | 24 |
| NM_153208 | *IQCK* | chr6:28462315-28593455 | 140 |
| NM_001012991 | *C16orf88* | chr6:28594515-28609132 | 12 |
| NM_020314 | *C16orf62* | chr6:28621208-28731692 | 130 |
| NM_014711 | *CP110* | chr6:28739342-28767554 | 21 |
| NM_016641 | *GDE1* | chr6:28767489-28787280 | 23 |
| NM_001105248 | *TMC5* | chr6:28792188-28836777 | 49 |
| NM_001160364 | *TMC7* | chr6:28898390-28938484 | 52 |
| NM_016138 | *COQ7* | chr6:28939812-28950941 | 9 |
| NM_001033380 | *Itpripl2* | chr6:28975917-28984544 | 9 |
| NM_016524 | *SYT17* | chr6:29031381-29103760 | 82 |
| NR_024436 | *LOC728276* | chr6:29109495-29127642 | 18 |
| NM_015092 | *SMG1* | chr6:29257813-29337314 | 74 |
| NM_015161 | *ARL6IP1* | chr6:29337969-29348302 | 9 |
| NM_001019 | *RPS15A* | chr6:29351515-29360443 | 5 |

**Table 4.3 List of primers used in dye-terminator sequencing.**
Follow-up sequencing was performed in additional samples for three variants. The PCR conditions used are described in *Materials and Methods*.

| SNP | Forward | Reverse |
|---|---|---|
| Chr6.24500625 | 5’-TGAGGGACTGGAACTGCTCT-3’ | 5’-AGTCCTGTGCGGAAATCTGA-3’ |
| Chr6.25681850 | 5’-TTTTGTTTGGCTGCCTTCTC-3’ | 5’-TGCCCACAGAAAAATCCCTA-3’ |
| Chr6.25714052 | 5’-GCCTTCCTCCCTTCTTCAGT-3’ | 5’-CGAAGGAGATGACACGGAGT-3’ |

**Table 4.4 Top 25 ranked findings from analysis for presbycusis in Border collies[a].** All top 25 hits from the EMMAX analysis reached statistical significance at the Bonferroni-corrected level, and all but one are on CFA6. All top 25 findings also reached genome-wide significance after empirical significance testing with permutation. Odds ratios calculated in PLINK demonstrate strong effects for all top hits.

| SNP | A1/A2 | Freq Cases | Freq Controls | $P_{EMMAX}$ | $P_{Allelic}$ | $P_{Perm}$ | OR (95% CI) |
|---|---|---|---|---|---|---|---|
| Chr6.25819273 | C/A | 0.00 | 0.78 | 1.09E-13 | 6.42E-14 | 1.00E-06 | N/A |
| Chr6.26517587 | A/G | 0.83 | 0.14 | 1.64E-10 | 2.71E-11 | 1.00E-06 | 28.29 (9.35 - 85.57) |
| Chr6.24591869 | T/C | 0.10 | 0.77 | 4.46E-10 | 1.09E-10 | 2.00E-06 | 0.034 (0.01 - 0.11) |
| Chr6.24577002 | C/T | 0.89 | 0.20 | 5.68E-10 | 2.91E-11 | 1.00E-06 | 34.77 (10.18 - 118.7) |
| Chr6.25174415 | C/G | 0.10 | 0.78 | 1.07E-09 | 8.06E-11 | 2.00E-06 | 0.03 (0.01 - 0.11) |
| Chr6.28753894 | A/G | 0.78 | 0.13 | 1.78E-09 | 1.36E-10 | 3.00E-06 | 24.11 (8.15 - 71.38) |
| Chr6.25181733 | G/A | 0.89 | 0.21 | 2.03E-09 | 9.38E-11 | 2.00E-06 | 31.17 (9.23 - 105.20) |
| Chr6.22844453 | G/A | 0.80 | 0.19 | 5.21E-09 | 3.07E-09 | 2.60E-05 | 17.6 (6.25 - 49.56) |
| Chr6.29363433 | T/G | 0.78 | 0.15 | 9.45E-09 | 1.07E-09 | 6.00E-06 | 19.81 (6.89 - 56.92) |
| Chr6.24570819 | T/G | 0.10 | 0.75 | 1.07E-08 | 3.28E-10 | 4.00E-06 | 0.037 (0.01 - 0.12) |
| Chr6.21475826 | C/T | 0.80 | 0.23 | 1.83E-08 | 3.87E-08 | 4.20E-04 | 13.23 (4.90 - 35.7) |
| Chr8.62484232 | T/G | 0.00 | 0.37 | 3.75E-08 | 1.44E-05 | 1.70E-01 | N/A |
| Chr6.25913101 | C/T | 0.00 | 0.61 | 4.14E-08 | 8.67E-10 | 5.00E-06 | N/A |
| Chr6.29470484 | T/C | 0.76 | 0.15 | 5.20E-08 | 3.15E-09 | 2.60E-05 | 18.53 (6.42 - 53.46) |
| Chr6.35491820 | G/C | 0.50 | 0.04 | 1.00E-07 | 2.68E-07 | 2.82E-03 | 25.00 (5.34 - 117.00) |
| Chr6.23160353 | C/A | 0.53 | 0.05 | 1.95E-07 | 1.45E-07 | 1.57E-03 | 19.53 (5.23 - 72.98) |
| Chr6.23166082 | G/A | 0.53 | 0.05 | 1.95E-07 | 1.45E-07 | 1.57E-03 | 19.53 (5.23 - 72.98) |
| Chr6.25900591 | G/A | 0.00 | 0.54 | 2.21E-07 | 2.37E-08 | 2.21E-04 | N/A |
| Chr6.26959216 | C/A | 0.03 | 0.61 | 2.40E-07 | 5.15E-09 | 3.60E-05 | 0.02 (0.002 - 0.13) |
| Chr6.34915222 | G/A | 0.63 | 0.14 | 3.21E-07 | 9.41E-07 | 1.18E-02 | 10.00 (3.74 - 26.77) |
| Chr6.23177930 | C/T | 0.53 | 0.06 | 3.48E-07 | 2.46E-07 | 2.57E-03 | 18.79 (5.02 - 70.3) |
| Chr6.26917473 | C/A | 0.03 | 0.59 | 3.67E-07 | 1.20E-08 | 9.50E-05 | 0.02 (0.002 - 0.14) |
| Chr6.24104844 | A/G | 0.08 | 0.66 | 3.76E-07 | 9.54E-09 | 7.30E-05 | 0.04 (0.01 - 0.15) |
| Chr6.34819558 | A/T | 0.03 | 0.38 | 8.32E-07 | 5.76E-05 | 5.31E-01 | 0.04 (0.005 - 0.33) |
| Chr6.22861769 | T/A | 0.50 | 0.05 | 1.58E-06 | 4.37E-07 | 4.99E-03 | 17.67 (4.73 - 66) |

[a]SNP: marker name (location information); A1: risk allele; A2: reference allele; Freq Cases: allele frequency of A1 in cases; Freq Controls: allele frequency of A1 in controls; $P_{EMMAX}$: p-values from EMMAX primary GWAS; $P_{Allelic}$: p-values from allelic association analysis; $P_{Perm}$: genome-wide (EMP2) permuted p-values from PLINK; OR (95% CI): odds ratio with 95% confidence interval as calculated with logistic regression.

**Table 4.5 NGS statistics.**
Each sample was run in a single lane for 76 sequencing cycles. Given the high number of variants called, we first filtered variants with regard to their genotype in cases and controls, filtering for variants called homozygous in the case sample and called not homozygous for that variant in either of the controls. We then focused on exonic and potentially functional non-coding variants, with priority given to top biological candidates.

| | **Control 1** | **Control 2** | **Case** |
|---|---|---|---|
| Total reads | 36,270,529 | 30,867,026 | 32,404,825 |
| Aligned reads | 33,564,330 | 27,899,663 | 30,252,775 |
| %Aligned reads | 92.5% | 90.4% | 93.4% |
| | | | |
| Mean bait coverage (X) | 905.9 | 658.2 | 785.3 |
| Mean target coverage (X) | 548.2 | 403.7 | 483.7 |
| Fold enrichment | 868.3 | 759.0 | 835.2 |
| | | | |
| %Target >2X | 77.6% | 78.7% | 81.6% |
| %Target >10X | 75.7% | 73.1% | 77.1% |
| %Target >20X | 73.1% | 68.6% | 73.8% |
| %Target >30X | 71.1% | 65.2% | 71.1% |
| | | | |
| Reads on target | 26,286,959 | 19,356,053 | 23,193,286 |
| % Aligned reads on target | 78% | 69% | 77% |

**Table 4.6 Gene coverage information by gene for each target capture sample.** (Col5: %Bases > 1X in Any Sample; Col6: %Bases >1X in Control 1; Col7: Control 1 average coverage (X); Col8: %Bases >1X in Control 2; Col9: Control 2 average coverage (X); Col10: %Bases >1X in Case; Col11: Case average coverage (X))

| Chr | Start | End | Gene | Col5 | Col6 | Col7 | Col8 | Col9 | Col10 | Col11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 23237568 | 23637888 | *HS3ST4* | 13% | 5% | 37 | 8% | 36 | 9% | 40 |
| 6 | 23977533 | 23996352 | *ZKSCAN2* | 100% | 97% | 728 | 99% | 539 | 100% | 691 |
| 6 | 24004232 | 24011663 | *AQP8* | 95% | 87% | 306 | 91% | 439 | 94% | 377 |
| 6 | 24047309 | 24085693 | *LCMT1* | 89% | 81% | 660 | 84% | 557 | 87% | 587 |
| 6 | 24058127 | 24059285 | *C8H9orf30* | 100% | 100% | 914 | 100% | 879 | 100% | 988 |
| 6 | 24058154 | 24058998 | *c9orf30* | 100% | 100% | 938 | 100% | 1008 | 100% | 1103 |
| 6 | 24100958 | 24195585 | *ARHGAP17* | 92% | 87% | 761 | 89% | 666 | 91% | 747 |
| 6 | 24202588 | 24236597 | *SLC5A11* | 92% | 82% | 617 | 87% | 529 | 90% | 613 |
| 6 | 24266253 | 24372846 | *TNRC6A* | 95% | 90% | 750 | 90% | 468 | 93% | 621 |
| 6 | 24499008 | 24530951 | *RBBP6* | 99% | 96% | 550 | 96% | 140 | 98% | 262 |
| 6 | 24645639 | 24731130 | *CACNG3* | 92% | 85% | 557 | 88% | 498 | 90% | 558 |
| 6 | 24761568 | 25085234 | *PRKCB* | 91% | 84% | 634 | 87% | 549 | 90% | 629 |
| 6 | 25149807 | 25152168 | *CHP2* | 100% | 100% | 1399 | 100% | 1391 | 100% | 1598 |
| 6 | 25175871 | 25189051 | *ERN2* | 91% | 86% | 654 | 90% | 853 | 90% | 720 |
| 6 | 25189004 | 25200121 | *PLK1* | 93% | 87% | 522 | 90% | 683 | 91% | 629 |
| 6 | 25209311 | 25233185 | *DCTN5* | 85% | 76% | 465 | 81% | 263 | 84% | 425 |
| 6 | 25233485 | 25261468 | *PALB2* | 91% | 80% | 477 | 81% | 238 | 87% | 346 |
| 6 | 25266061 | 25275929 | *NDUFAB1* | 99% | 94% | 773 | 97% | 570 | 98% | 737 |
| 6 | 25283922 | 25299748 | *UBFD1* | 97% | 92% | 618 | 96% | 614 | 97% | 643 |
| 6 | 25299942 | 25321502 | *EARS2* | 94% | 86% | 582 | 91% | 606 | 93% | 588 |
| 6 | 25341027 | 25368230 | *GGA2* | 88% | 83% | 821 | 83% | 668 | 86% | 736 |
| 6 | 25379018 | 25459052 | *COG7* | 83% | 75% | 596 | 75% | 500 | 80% | 582 |
| 6 | 25463920 | 25486946 | *SCNN1B* | 89% | 79% | 437 | 86% | 562 | 87% | 446 |
| 6 | 25594643 | 25617204 | *SCNN1G* | 97% | 87% | 693 | 94% | 673 | 95% | 713 |
| 6 | 25650946 | 25721762 | *USP31* | 94% | 90% | 761 | 90% | 407 | 93% | 556 |
| 6 | 25843382 | 25934014 | *HS3ST2* | 93% | 86% | 573 | 89% | 586 | 91% | 601 |
| 6 | 26135411 | 26196802 | *OTOA* | 91% | 82% | 544 | 87% | 424 | 88% | 496 |
| 6 | 26209563 | 26244138 | *METTL9* | 93% | 88% | 496 | 87% | 208 | 92% | 336 |
| 6 | 26217896 | 26222970 | *Igsf6* | 100% | 100% | 708 | 97% | 322 | 100% | 386 |
| 6 | 26325472 | 26344913 | *CDR2* | 88% | 79% | 350 | 82% | 168 | 86% | 278 |
| 6 | 26357857 | 26386924 | *POLR3E* | 93% | 86% | 657 | 91% | 842 | 91% | 743 |
| 6 | 26397174 | 26462173 | *EEF2K* | 87% | 74% | 436 | 82% | 449 | 84% | 498 |
| 6 | 26501885 | 26561226 | *VWA3A* | 90% | 80% | 623 | 84% | 569 | 89% | 647 |
| 6 | 26569920 | 26636167 | *C16orf52* | 83% | 76% | 273 | 74% | 70 | 79% | 145 |
| 6 | 26643205 | 26655344 | *PDZD9* | 99% | 95% | 721 | 95% | 475 | 99% | 616 |
| 6 | 26656187 | 26681978 | *UQCRC2* | 90% | 85% | 591 | 83% | 286 | 86% | 453 |
| 6 | 26704127 | 26834863 | *Abca14* | 43% | 37% | 135 | 35% | 34 | 37% | 57 |
| 6 | 26888834 | 27194268 | *Abca16* | 45% | 38% | 165 | 38% | 46 | 41% | 77 |
| 6 | 27214760 | 27231848 | *CRYM* | 99% | 96% | 680 | 98% | 471 | 99% | 621 |
| 6 | 27237043 | 27245352 | *ANKS4B* | 95% | 92% | 828 | 90% | 434 | 91% | 618 |
| 6 | 27257838 | 27268770 | *ZP2* | 100% | 100% | 1154 | 99% | 495 | 100% | 871 |
| 6 | 27283092 | 27293013 | *TMEM159* | 92% | 89% | 663 | 86% | 384 | 89% | 516 |
| 6 | 27305543 | 27465095 | *DNAH3* | 84% | 77% | 560 | 77% | 336 | 81% | 456 |
| 6 | 27468260 | 27488430 | *LYRM1* | 99% | 93% | 710 | 96% | 441 | 97% | 554 |
| 6 | 27488425 | 27524621 | *DCUN1D3* | 97% | 93% | 709 | 95% | 430 | 96% | 581 |
| 6 | 27533246 | 27577425 | *LOC81691* | 69% | 65% | 312 | 64% | 144 | 67% | 229 |
| 6 | 27577410 | 27586378 | *ERI2* | 100% | 95% | 416 | 93% | 88 | 99% | 163 |
| 6 | 27585660 | 27604071 | *ACSM3* | 100% | 96% | 573 | 93% | 182 | 98% | 370 |
| 6 | 27634650 | 27643499 | *THUMPD1* | 80% | 75% | 340 | 73% | 114 | 75% | 209 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 27644513 | 27670891 | *Acsm4* | 94% | 86% | 409 | 85% | 173 | 92% | 278 |
| 6 | 27684924 | 27686135 | *nono* | 100% | 100% | 3034 | 100% | 2732 | 100% | 2612 |
| 6 | 27688755 | 27689305 | *LOC728643* | 100% | 100% | 2300 | 100% | 2764 | 100% | 2370 |
| 6 | 27893180 | 27920107 | *ACSM2A* | 77% | 71% | 600 | 71% | 333 | 74% | 470 |
| 6 | 27935924 | 27966023 | *ACSM5* | 66% | 60% | 427 | 62% | 319 | 64% | 376 |
| 6 | 27971602 | 28013520 | *PDILT* | 91% | 83% | 588 | 84% | 405 | 88% | 496 |
| 6 | 28019265 | 28034478 | *UMOD* | 98% | 93% | 768 | 95% | 645 | 98% | 708 |
| 6 | 28039975 | 28054408 | *GP2* | 98% | 91% | 800 | 91% | 663 | 97% | 690 |
| 6 | 28277437 | 28317406 | *GPR139* | 89% | 84% | 545 | 83% | 431 | 87% | 475 |
| 6 | 28440878 | 28461832 | *GPRC5B* | 97% | 88% | 614 | 94% | 643 | 95% | 643 |
| 6 | 28463315 | 28592455 | *IQCK* | 84% | 75% | 430 | 75% | 237 | 81% | 349 |
| 6 | 28595515 | 28608132 | *C16orf88* | 75% | 73% | 591 | 73% | 543 | 75% | 560 |
| 6 | 28622208 | 28730692 | *C16orf62* | 94% | 86% | 558 | 90% | 452 | 93% | 547 |
| 6 | 28740342 | 28766554 | *CP110* | 92% | 89% | 493 | 86% | 135 | 89% | 281 |
| 6 | 28768489 | 28786280 | *GDE1* | 95% | 88% | 393 | 85% | 182 | 91% | 312 |
| 6 | 28793188 | 28835777 | *TMC5* | 90% | 82% | 515 | 85% | 339 | 88% | 476 |
| 6 | 28899390 | 28937484 | *TMC7* | 87% | 77% | 385 | 80% | 306 | 83% | 344 |
| 6 | 28940812 | 28949941 | *COQ7* | 86% | 74% | 565 | 80% | 280 | 83% | 443 |
| 6 | 28976917 | 28983544 | *Itpripl2* | 92% | 88% | 501 | 92% | 366 | 90% | 491 |
| 6 | 29032381 | 29102760 | *SYT17* | 89% | 78% | 480 | 84% | 436 | 88% | 516 |
| 6 | 29110495 | 29126642 | *LOC728276* | 88% | 77% | 528 | 85% | 571 | 87% | 554 |
| 6 | 29258813 | 29336314 | *SMG1* | 94% | 89% | 496 | 87% | 172 | 92% | 312 |
| 6 | 29338969 | 29347302 | *ARL6IP1* | 100% | 99% | 495 | 99% | 213 | 100% | 276 |
| 6 | 29352515 | 29359443 | *RPS15A* | 100% | 100% | 959 | 99% | 553 | 100% | 695 |

**Table 4.7 Summary of variants homozygous in cases and not in controls using ANNOVAR.**

| | In target | In target & homozygous in case but not controls |
|---|---|---|
| Downstream | 86 | 23 |
| Exonic | 106 | 26 |
| Synonymous | 67 | 19 |
| Nonsynonymous | 38 | 7 |
| Stopgain | 1 | 0 |
| Intergenic | 953 | 140 |
| Intronic | 3538 | 718 |
| Splicing | 2 | 0 |
| Upstream | 96 | 17 |
| 3' UTR | 10 | 2 |
| 5' UTR | 28 | 9 |

*3’ UTR: untranslated 3’; 5’ UTR: untranslated 5’*

**Table 4.8 Exonic variants for deafness on CFA6.**
A list of the 26 exonic variants for CFA6 plus annotations is given in Supplemental Table 4. Gene annotations and predicted amino acid (AA) changes (single letter AA abbreviations flanking AA position) are given with reference to the gene in the human unless the gene is not present in human, in which case it is given for the species noted (Mus – mouse, Sac – yeast, Bos – cow, Rat – rat). Non-synonymous SNPs (nsSNP) are marked in bold. In addition to the called genotypes for each sample, the sequence coverage for that SNP is also provided. Finally, the phastCons4Way score provides a measure of conservation for each position, where values closer to 1 indicate the base is more highly conserved across species. Conservation is based on alignment with human (hg17), mouse (mm6), and rat (rn3). CFA: canine chromosome; Position: base position; Ref: reference allele from genome; Alt: alternate allele observed in sample[s]; Call: 0 = reference allele, 1 = alternate allele; Cov: coverage (X); phastCons score: phastCons4Way score from UCSC genome browser.

Of the 26 putative exonic variants, only 8 were annotated to be non-synonymous changes. Four nsSNPs were found in *Abca14*, which was the gene with the most nsSNPs. *Abca14* is an ATP binding cassette transporter gene that has only been annotated in the genomes of rodents [38]. Conservation scores for all four of these nsSNPs were low, suggesting that this gene may not be active and thus tolerant of non-synonymous changes more readily. There was an additional gene containing an nsSNPs that is not readily linked to hearing function or expression (*EEF2K*).

| Type | Locus | CFA | Position | Ref | Alt | Control 1 | | Control 2 | | Case | | phastCons |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Call | Cov | Call | Cov | Call | Cov | Score |
| **nsSNP** | ***RBBP6*, exon18, p.T1397N** | **6** | **2450062** | **G** | **T** | **0/0** | **249** | **0/0** | **72** | **1/1** | **226** | **0.001** |
| ssSNP | *RBBP6*, exon11, p.E445E | 6 | 2450847 | T | C | 1/1 | 248 | 1/1 | 123 | 0/0 | 226 | 0.925 |
| ssSNP | *ERN2*, exon9, p.L260L | 6 | 2517874 | T | C | 0/0 | 236 | 0/1 | 241 | 1/1 | 240 | 0.871 |
| ssSNP | *ERN2*, exon19, p.F578F | 6 | 2518589 | C | T | 1/1 | 233 | 0/1 | 223 | 0/0 | 214 | 0.925 |
| ssSNP | *ERN2*, exon25, p.D838D | 6 | 2518830 | C | T | 1/1 | 208 | 0/1 | 245 | 0/0 | 244 | 0.949 |
| ssSNP | *PLK1*, exon9, p.L511L | 6 | 2518984 | A | G | 1/1 | 233 | 0/1 | 236 | 0/0 | 244 | 0.792 |
| ssSNP | *PLK1*, exon9, p.E488E | 6 | 2518990 | T | C | 1/1 | 242 | 0/1 | 247 | 0/0 | 249 | 0.831 |
| ssSNP | *PLK1*, exon1, p.K97K | 6 | 2519981 | C | T | 1/1 | 238 | 0/1 | 233 | 0/0 | 241 | 0.971 |
| **nsSNP** | ***USP31*, exon17, p.I847V** | **6** | **2571405** | **A** | **G** | **0/0** | **245** | **0/1** | **246** | **1/1** | **245** | **0.950** |
| **nsSNP** | ***EEF2K*, exon1, p.N62K** | **6** | **2644265** | **G** | **T** | **1/1** | **249** | **0/1** | **250** | **0/0** | **249** | **0.627** |
| **nsSNP** | ***Abca14* (Mus), exon26, p.M1292L** | **6** | **2690986** | **T** | **G** | **0/0** | **53** | **0/1** | **18** | **1/1** | **16** | **0.058** |
| **nsSNP** | ***Abca14* (Mus), exon23, p.L1134I** | **6** | **2692454** | **G** | **T** | **0/0** | **230** | **0/1** | **63** | **1/1** | **189** | **0.013** |
| **nsSNP** | ***Abca14* (Mus), exon16, p.V699I** | **6** | **2695157** | **C** | **T** | **0/0** | **243** | **0/1** | **247** | **1/1** | **247** | **0.045** |
| **nsSNP** | ***Abca14* (Mus), exon9, p.I472M** | **6** | **2697267** | **T** | **C** | **0/0** | **220** | **0/1** | **114** | **1/1** | **242** | **0.145** |
| ssSNP | *DNAH3*, exon56, p.R2825R | 6 | 2743688 | G | A | 0/0 | 247 | 0/1 | 246 | 1/1 | 246 | 0.980 |
| ssSNP | *DNAH3*, exon60, p.D3733D | 6 | 2745670 | C | T | 0/0 | 243 | 0/1 | 243 | 1/1 | 245 | 0.031 |
| ssSNP | *LOC57020*, exon11, p.D462D | 6 | 2789848 | G | A | 0/0 | 244 | 0/0 | 246 | 1/1 | 242 | 0.352 |
| ssSNP | *C16orf62*, exon26, p.T707T | 6 | 2866188 | C | T | 0/0 | 249 | 0/0 | 67 | 1/1 | 245 | 0.972 |
| ssSNP | *C16orf62*, exon9, p.S232S | 6 | 2870954 | A | G | 0/0 | 243 | 0/0 | 226 | 1/1 | 246 | 0.972 |
| ssSNP | *CP110*, exon3, p.K201K | 6 | 2875037 | C | T | 0/0 | 191 | 0/0 | 91 | 1/1 | 221 | 0.463 |
| ssSNP | *CP110*, exon1, p.L142L | 6 | 2875351 | T | C | 1/1 | 248 | 0/1 | 148 | 0/0 | 230 | 0.918 |
| ssSNP | *CP110*, exon1, p.T17T | 6 | 2875389 | G | A | 0/0 | 247 | 0/0 | 222 | 1/1 | 246 | 0.518 |
| ssSNP | *COQ7* (Sac), exon4, p.K131K | 6 | 2894844 | A | G | 0/0 | 240 | 0/0 | 249 | 1/1 | 249 | 0.923 |
| ssSNP | *ITPRIPL2*, exon1, p.A365A | 6 | 2897845 | C | T | 0/0 | 55 | 0/0 | 149 | 1/1 | 88 | 0.880 |
| ssSNP | *ITPRIPL2*, exon1, p.L414L | 6 | 2897860 | T | C | 0/0 | 80 | 0/0 | 165 | 1/1 | 102 | 0.141 |
| ssSNP | *ITPRIPL2*, exon1, p.L415L | 6 | 2897860 | T | C | 0/0 | 77 | 0/0 | 156 | 1/1 | 104 | 0.063 |

**Table 4.9 Summary of association results with combined data from all stages of study[a].**

The strength of association of the three variants is shown. Chr6.24500625 in *RBBP6* and Chr6.25681850 in *USP31* remain highly significant after inclusion of more cases and controls.

| | | | Primary | | | Replication | | | |
|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Gene** | **A1/A2** | **$Fr_{Case}/Fr_{Cont}$<br>n=23/n=101** | **P** | **OR** | **$Fr_{Case}/Fr_{Cont}$<br>n=16/n=265** | **P** | **OR** | **Comb. P** |
| Chr6.24500625 | *RBBP6* | T/G | 0.91/0.38 | 1.98E–10 | 16.32<br>(5.62-47.41) | 0.69/0.41 | 0.0019 | 3.20<br>(1.49-6.89) | 1.01E–9 |
| Chr6.25681850 | *USP31* | G/T | 0.98/0.31 | 3.24E–16 | 97.98<br>(13.19-727.90) | 0.72/0.23 | 8.57E–10 | 8.55<br>(3.85-18.96) | 6.16E–22 |
| Chr6.25714052 | *USP31* | G/A | 0.98/0.74 | 0.0012 | 13.52<br>(1.81-100.90) | 0.91/0.71 | 0.016 | 3.96<br>(1.19-13.19) | 0.00015 |

[a] SNP: marker name (location information); A1: risk allele; A2: reference allele; $Fr_{Case}$: allele frequency of A1 in cases;
$Fr_{Cont}$: allele frequency of A1 in controls; P: p-values from allelic association analysis; OR: odds ratio with 95% confidence interval; Comb. P: combined p-value from meta-analysis.

**Figure 2.1 Frequency of mtDNA singleton variants among haplogroup N pancreatic cancer cases and controls, San Francisco Bay Area, California (1995-1999).**
Fisher's exact test: p-value<0.05*, p-value=0.01**

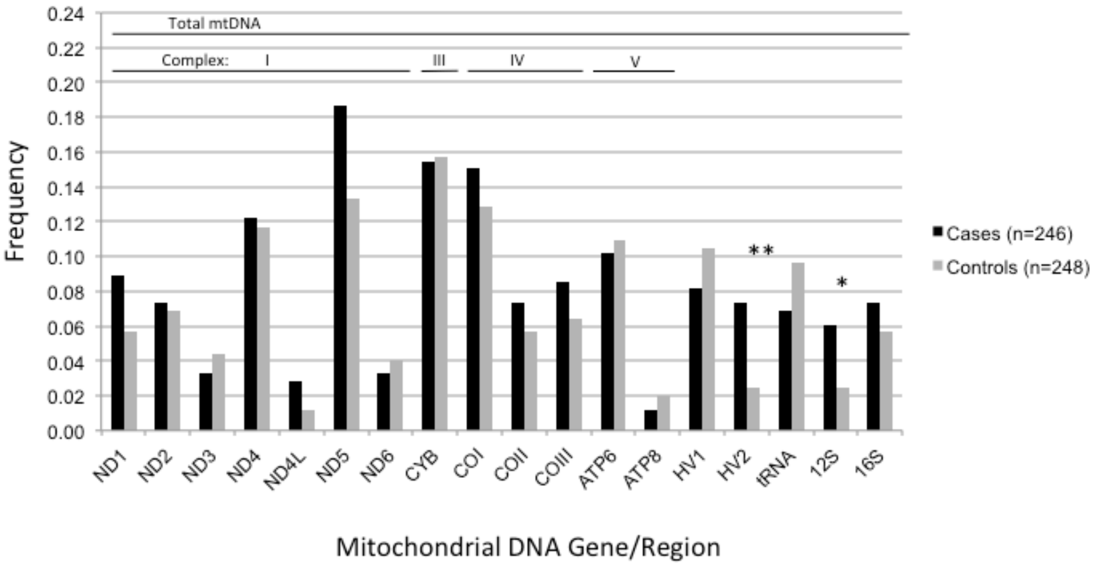


**Figure 2.2 Frequency of mtDNA singleton variants among haplogroup L pancreatic cancer cases and controls, San Francisco Bay Area, California (1995-1999).**
Fisher's exact test: p-value<0.05*, p-value=0.01**

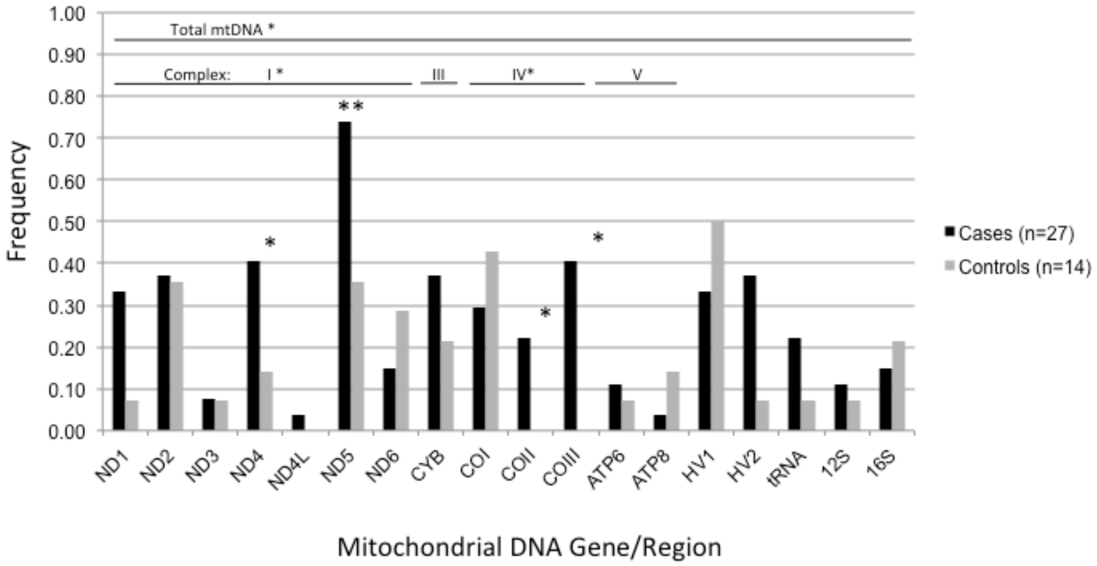


**Figure 2.3 Frequency of mtDNA singleton variants among haplogroup M pancreatic cancer cases and controls, San Francisco Bay Area, California (1995-1999).**
Fisher's exact test: p-value<0.05*

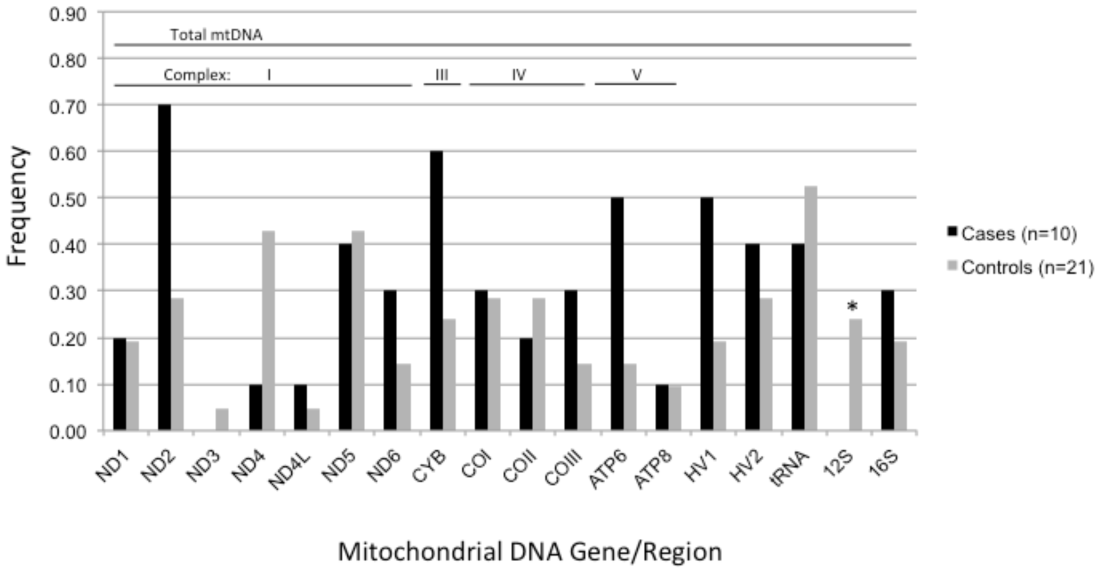
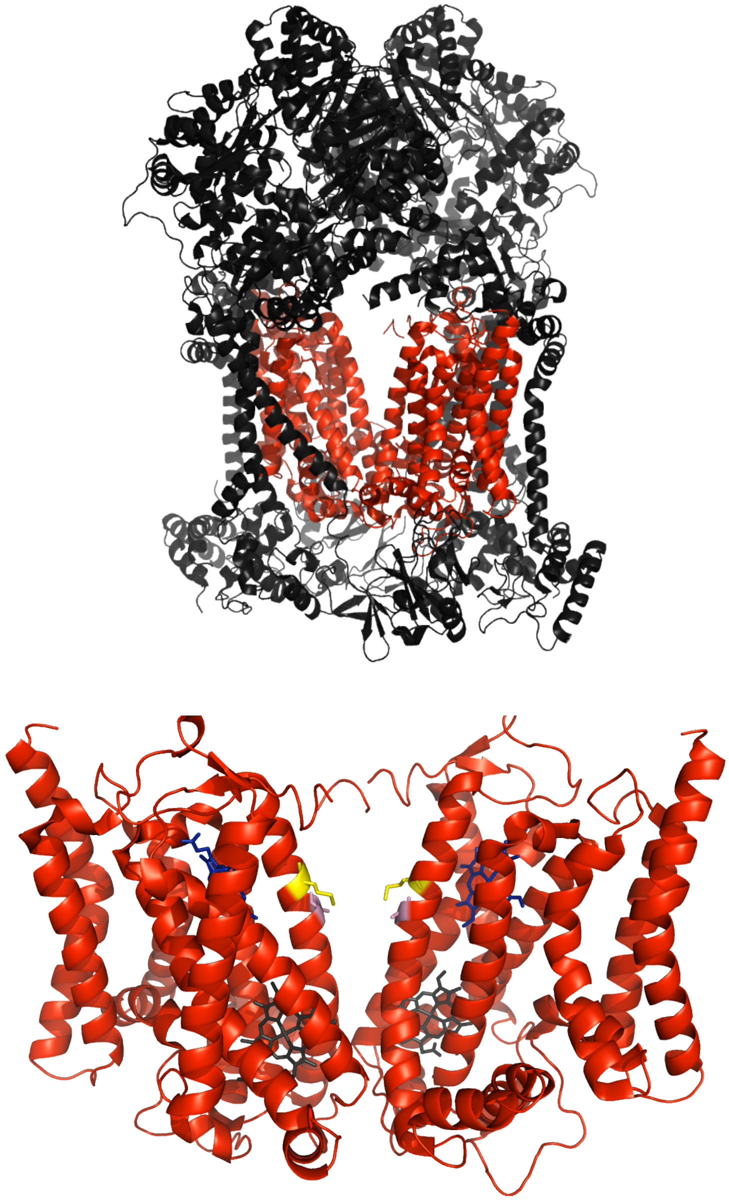


**Figure 2.4 Structure of the dimeric bovine cytochrome bc1 complex at 2.28 Å resolution (PDB2a06).**

A) Cytochrome bc1 complex with mtDNA-encoded *CytB* (red) and nDNA-encoded subunits (gray) indicated.

B) Close-up of *CytB* dimer indicating p.A191T (Purple) and p.T194M (yellow) positions located in the Qi binding pocket of complex III, where quinone is reduced by *CytB*. The $b_L$ heme (blue) adjacent to the Qo site and $b_H$ heme (grey) adjacent to the Qi site are also indicated. The T194M variant occurs at a residue that undergoes significant conformational changes upon contact with antimycin A, a pharmacological inhibitor of the Qi site.  In the presence of antimycin A, complex III produces high quantities of superoxide indicating that inhibition at this site blocks electron transfer from cytochrome b to quinone causing a buildup of semiquinone resulting in increased ROS production

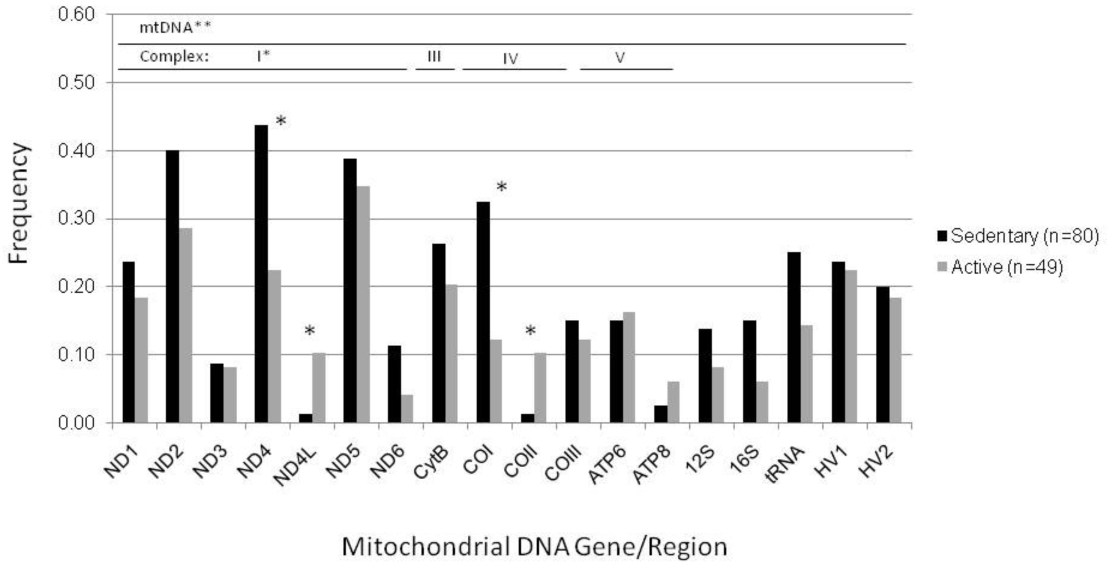
A)



B)

**Figure 2.5 Frequency of mtDNA singleton variants unique to sedentary or active Health ABC Study participants.**

Sedentary vs. Active, Fisher's Exact Test P-value <0.05*, <0.01**.



[a] Sedentary, physical activity level <1.7
[b] Active, physical activity level ≥1.7

**Figure 3.1 Mean relative cell viability measurements.**
The calculation was done by dividing final cell count by initial cell count) for agents representative of drug classes studied.

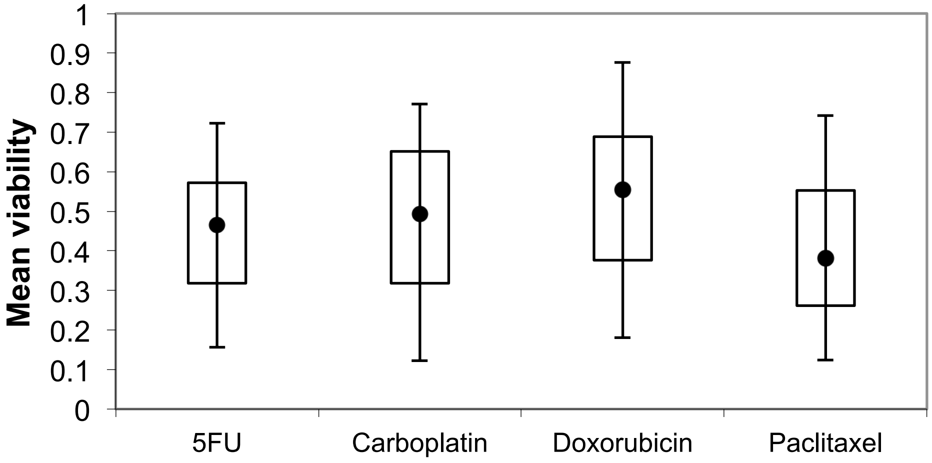


**Figure 3.2 Venn diagram illustrating the overlap between drug pathway.** Genes were selected from five drug pathways. There is some overlap between pathways, but because each drug class has its unique mechanism of action, the pathways do not typical share the same drug targets. However, some transporter and drug metabolizing proteins are shared between pathways.

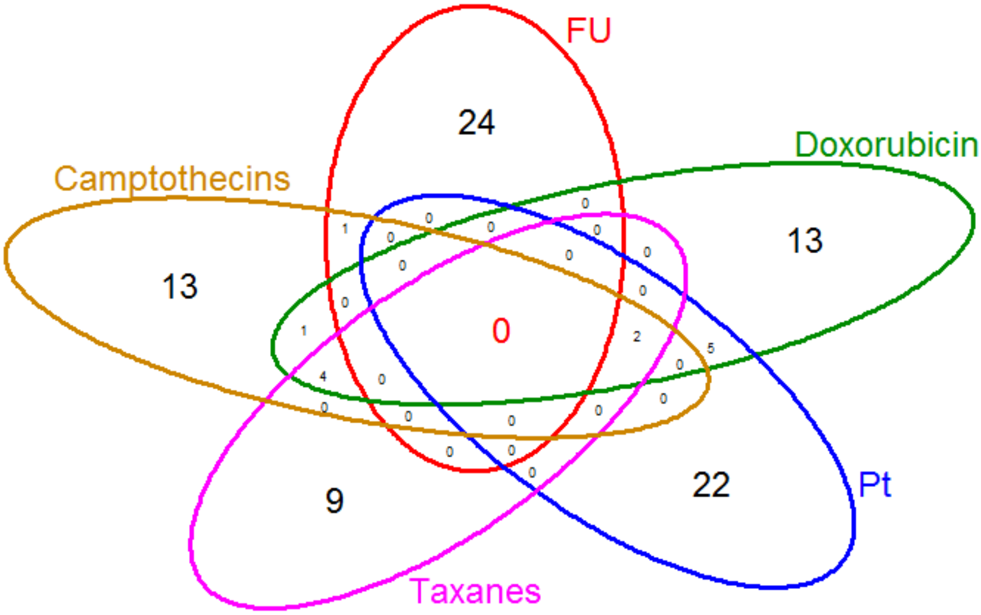


**Figure 3.3 Primer design and preparation process.**
The primer pairs were designed using an iterative pipeline developed by RainDance Technologies. The primers were then synthesized, put into droplets, and pooled together.

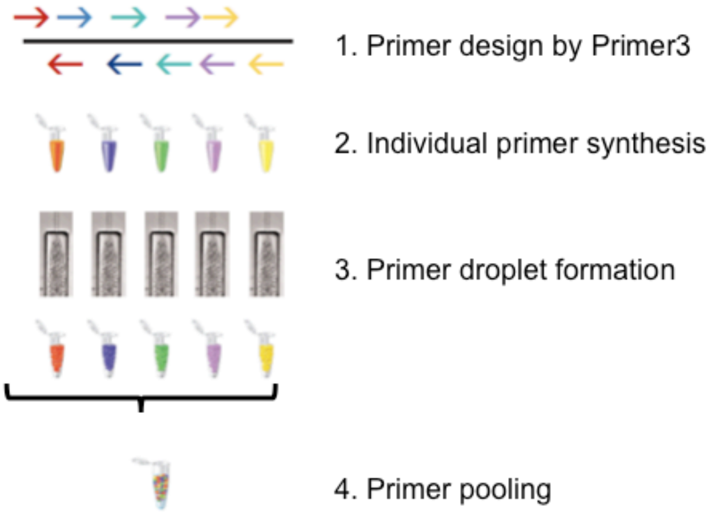

1. Primer design by Primer3
2. Individual primer synthesis
3. Primer droplet formation
4. Primer pooling

**Figure 3.4 Droplet merging with DNA template.**
The DNA template was mixed with the primer droplets to create complete PCR reactions. After thermal cycling, the PCR products were subjected to sample preparation for next generation sequencing.

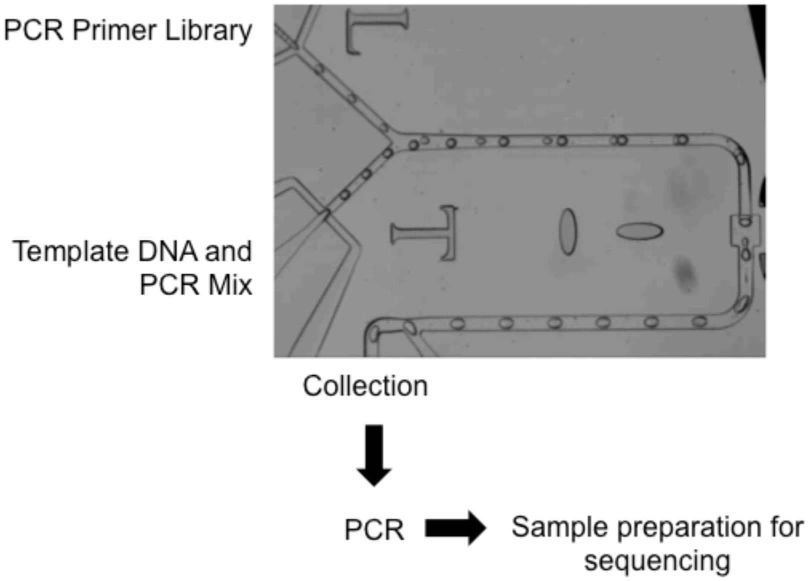


**Figure 3.5 Alignment statistics.**
“Total reads” represents all of the reads that were collected for a given sample regardless of the source of the data. Only unique alignments were considered.

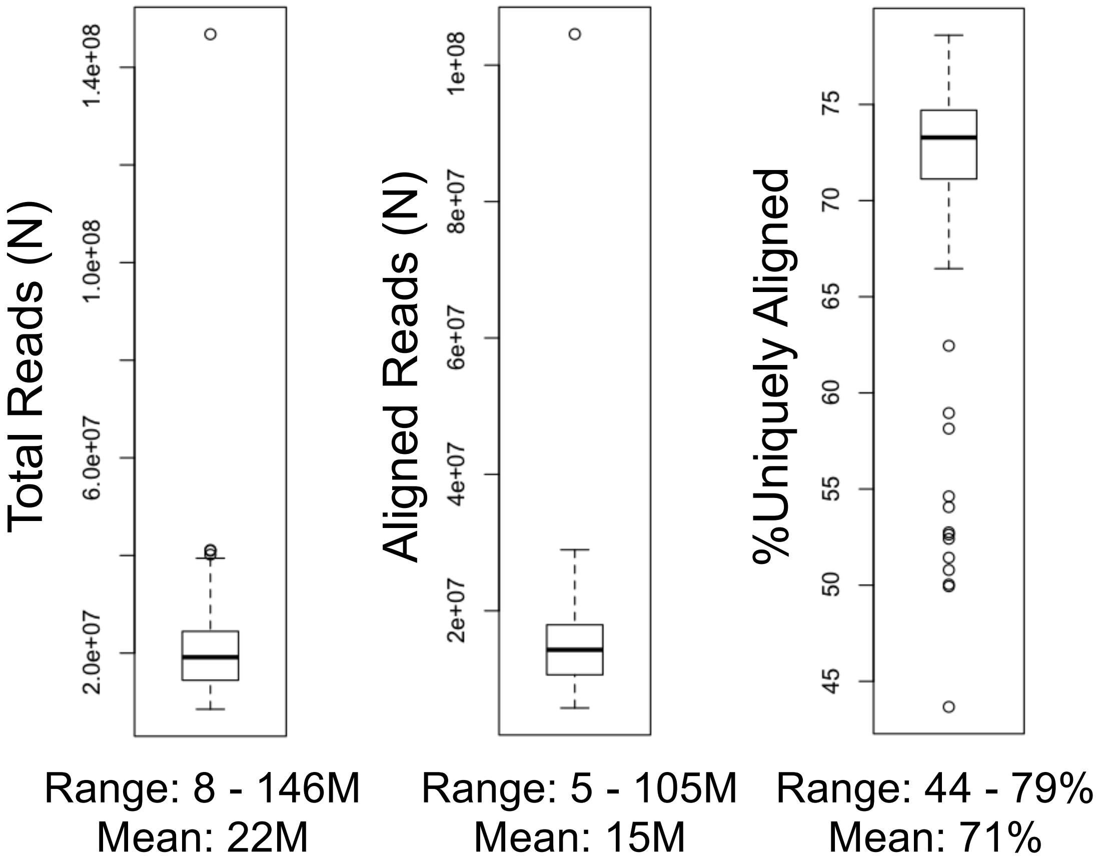


**Figure 3.6 Coverage distribution.**
Coverage distribution is shown for one of the tiled regions. The primer sites are marked in black. As shown, regions with high coverage coincide with the primer sites.

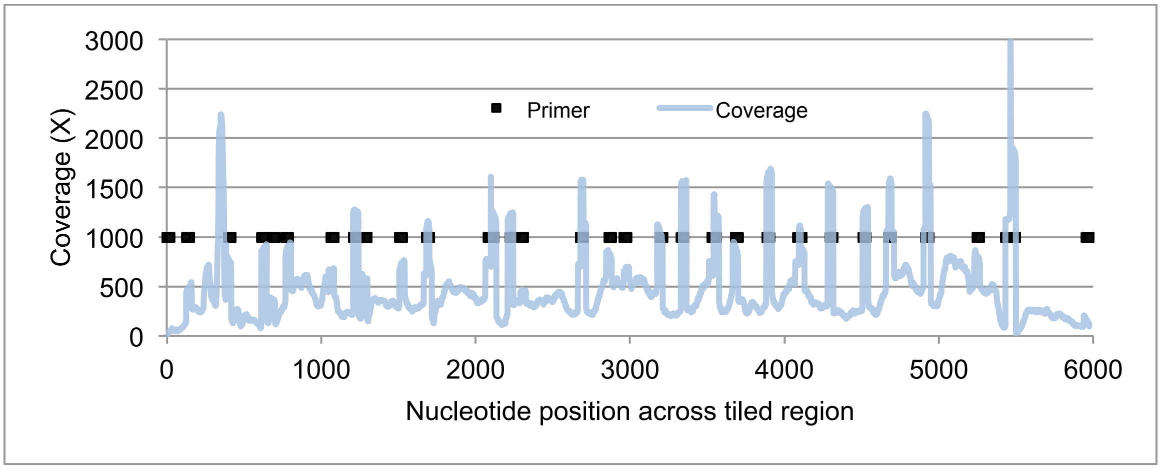


**Figure 4.1 Manhattan plot of GWAS for adult-onset deafness.**
Chromosome markers are plotted on the x-axis in order and alternately shaded. The -$\log_{10}$(p-value) is plotted on the y-axis. The red line indicates significance at the Bonferroni-corrected level for 30,000 SNPs. There is extensive regional support for an association on CFA6. The inset shows an enlargement of the 25-Mb association region, including genes of interest. The raw GWAS and permuted p-values ($P_{\text{perm-gw}}$) for the top SNP are also given.

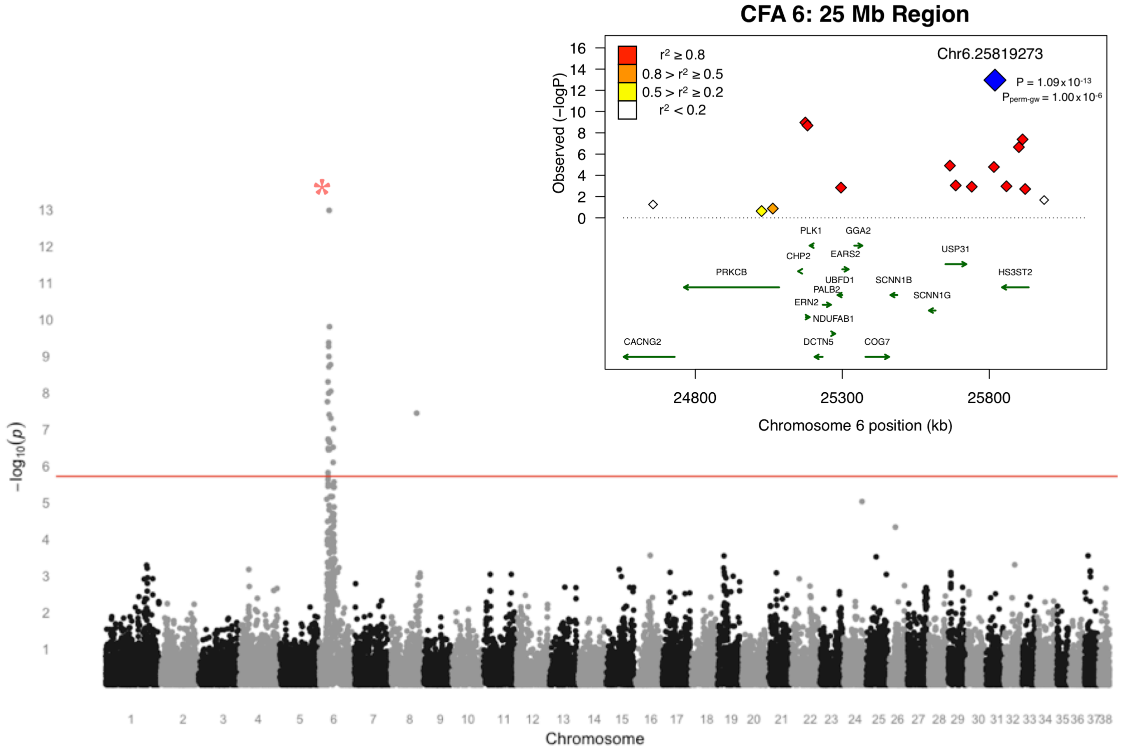


**Figure 4.2 Haplotypes in CFA6 region 25 Mb.**
Each box color represents a different genotype, as indicated by the key; dogs are listed in rows and SNPs in columns. Case dogs are all homozygous for a single haplotype spanning 7 markers, and all but one case also share an 11-SNP haplotype (for which the single dog is heterozygous). One sample used as a control (marked with *) also carries the 11-SNP risk haplotype.

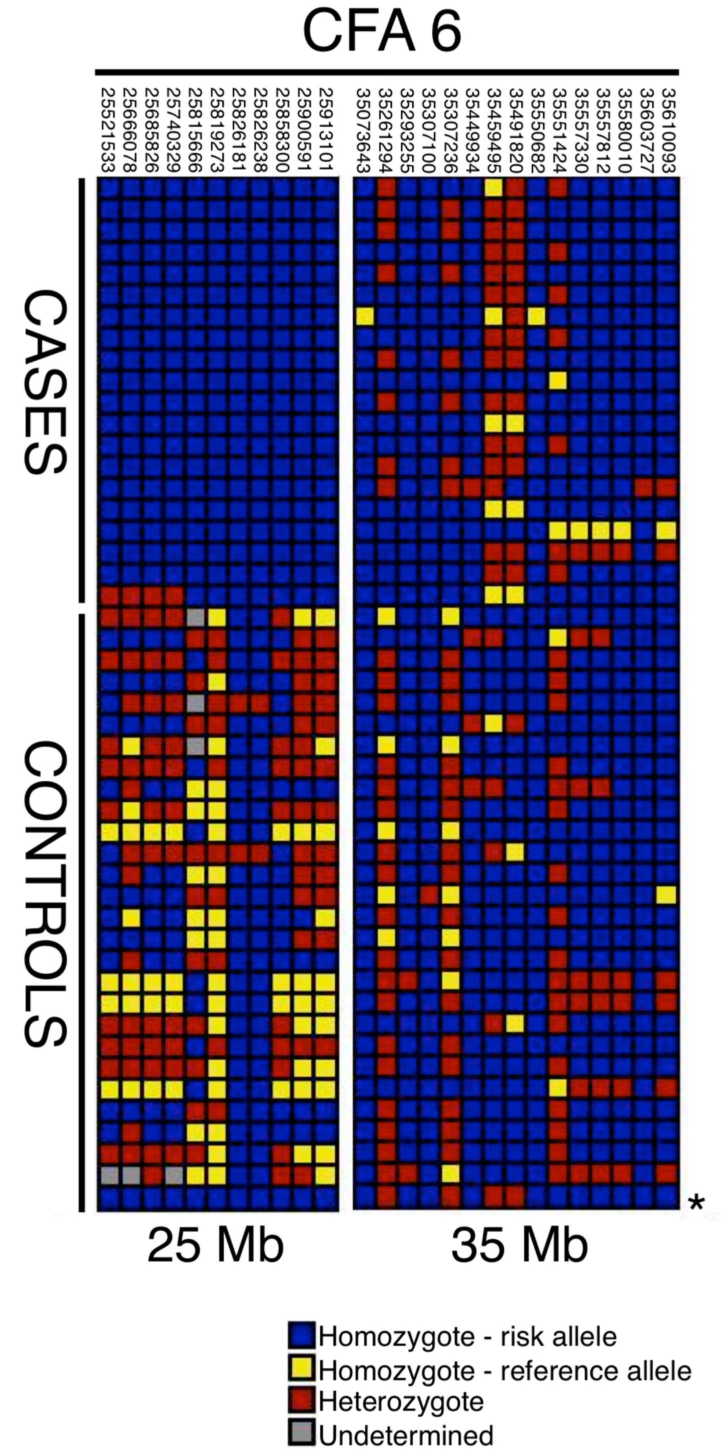


**Figure 4.3 Number of gaps in the assembly from 23 Mb to 29 Mb by size.**
There were 80 gaps in the canFam2 assembly ranging from 1 bp to 2707 bp (mean = 388 bp; median = 217 bp). The gaps sum to 31089 bp, or about 0.5% of the sequence within the region.

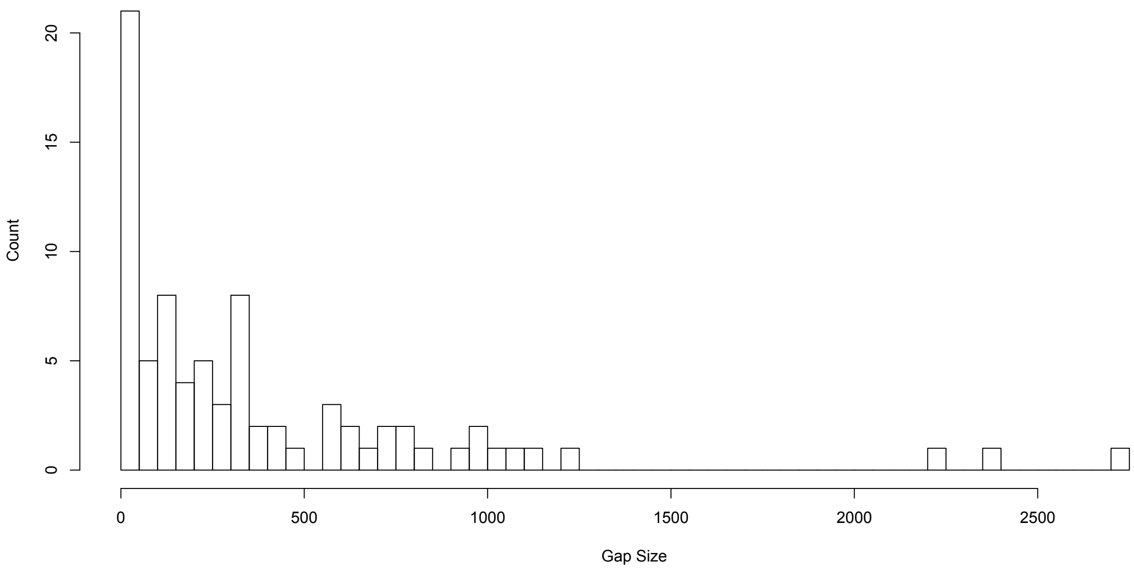


**Figure 4.4 Sequencing quality scores of target capture region.**
As shown, most (98%) of the bases have a scores equal or bigger than 40 (deemed high confidence).

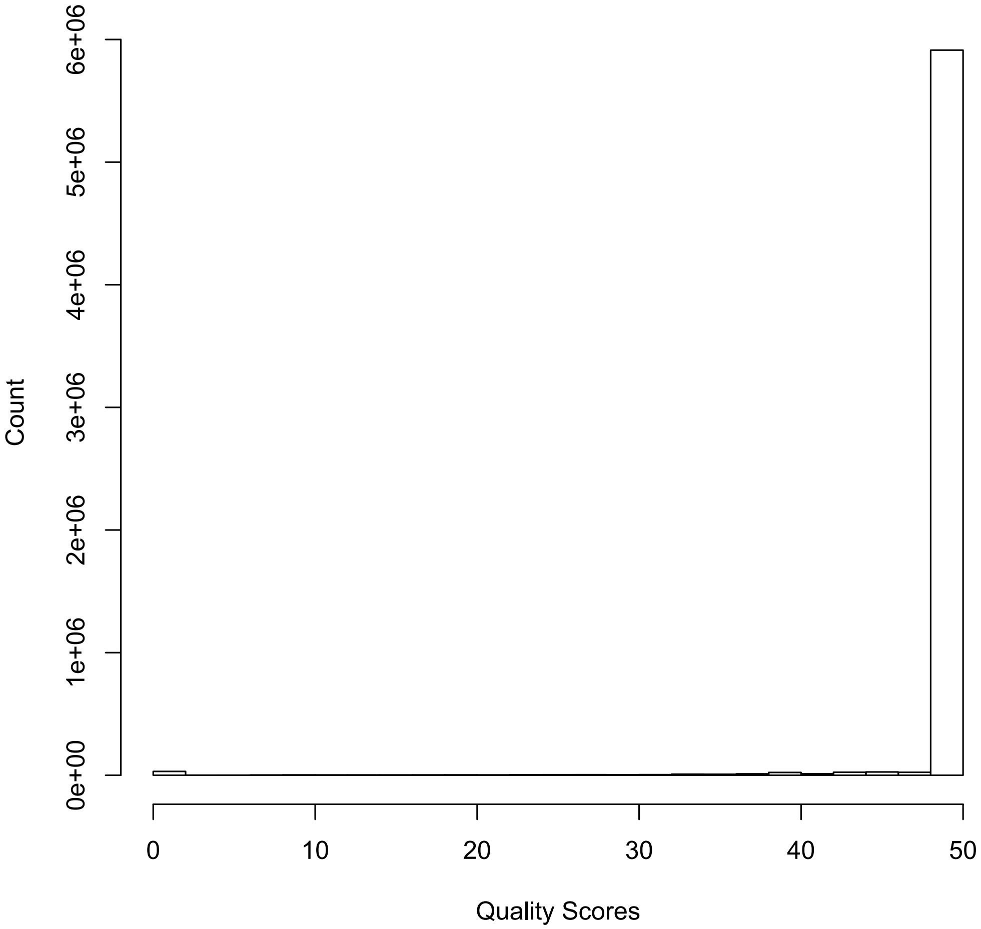


**Figure 5.1 Use of nanochannel arrays.**
(**a**) In a microfluidic environment, long (>100 kb) DNA fragments are in the coiled ball form and clog the entrance of the nanochannel array, as it is energetically unfavorable for the molecules to uncoil and enter the nanochannels. (**b**) A gradient region is placed in front of the nanochannels. Here, the physical confinement is sufficiently dense that the molecules are forced to flow by the pillars, where they uncoil and stream into the nanochannels without clogging. (**c**) Fabrication of the nanochannel array using interference lithography to produce 120 nm channels in silicon followed by tuning to a smaller diameter with material deposition and capping with a glass cover to allow for fluorescence imaging. (**d**) A profile scanning electron microscopy (SEM) image of 45 nm channels. (**e**) An SEM image of the 45 nm channels patterned on the silicon substrate before bonding to the glass.

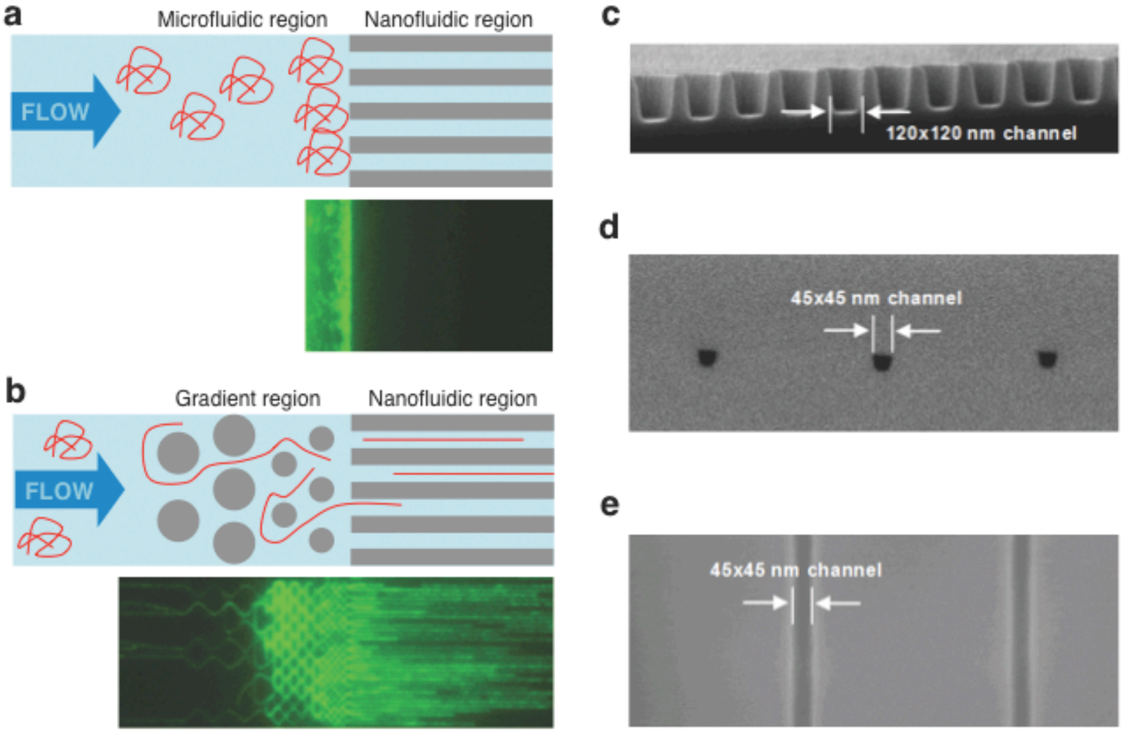


**Figure 5.2 Nano-mapping.**
(**a**) Nick labeling by Nt.BspQI and DNA polymerase is accomplished by top strand DNA cleavage (blue arrow) one nucleotide 3' from the recognition sequence (in bold italics) followed by incorporation of fluorescent nucleotide analogs (in red) with concomitant DNA strand displacement. (**b**) The DNA molecule is stained with YOYO-1 and loaded into the port of a nanoarray flowcell (left panel). The DNA molecules are introduced into the region with pillars and micron-scale relaxation channels by an electric field where they unwind and linearize (top right panel). Finally, the DNA molecules are pushed by a low-voltage electrical pulse, and they enter the 45 nm nanochannels, where they are stretched uniformly to 85% of the length of perfectly linear B-DNA (bottom right panel). The DNA is visualized as blue linear structures in the nanochannels, with green labels marking the Nt.BspQI nick sites. (**c**) The length of the DNA molecules and the positions of nick labels on each DNA molecule are determined after automated image capture. The fragment size profile of a 183 kb BAC is shown, with the narrow peak width indicating uniform DNA linearization. (**d**) The DNA molecules are clustered into groups (representing individual BACs) based on nick labeling pattern similarity. As BAC molecules can enter the nanochannels in either orientation, each BAC is represented by two clusters with opposite orientations (top panel). After combining the two clusters, histogram plots of nick-labels (bottom panel) are used to define the locations of Nt.BspQI sites with good precision.

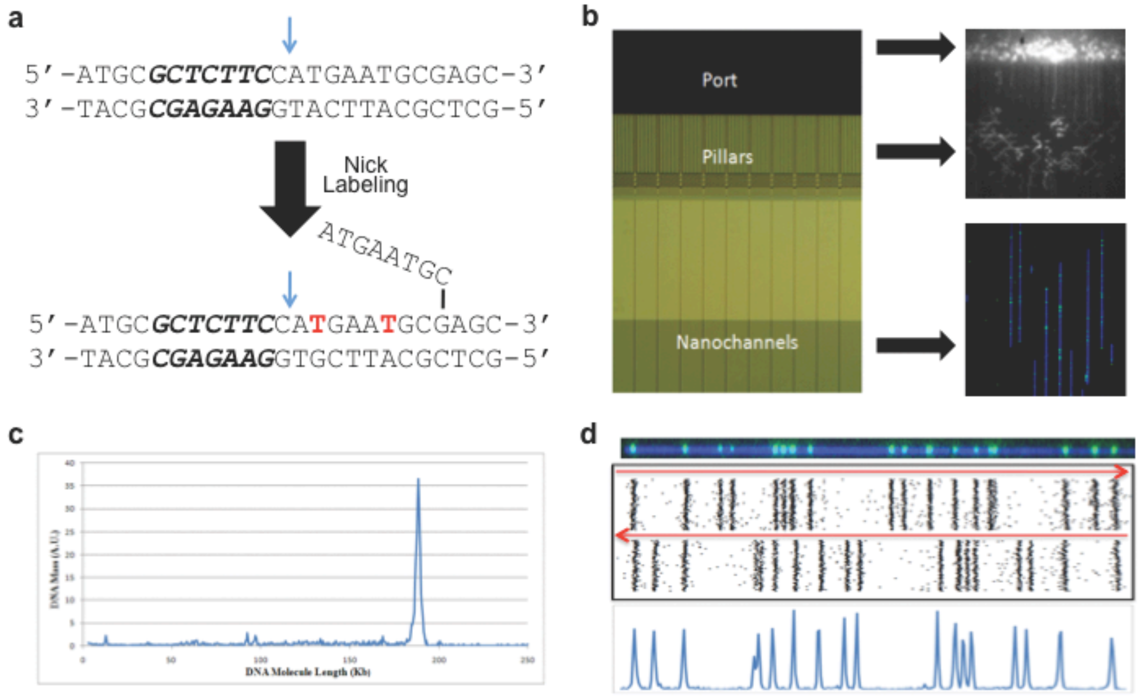

a
5'-ATGC***GCTCTTC***CATGAATGCGAGC-3'
3'-TACG***CGAGAAG***GTACTTACGCTCG-5'
Nick Labeling
ATGAATGC
5'-ATGC***GCTCTTC***CATGAATGCGAGC-3'
3'-TACG***CGAGAAG***GTGCTTACGCTCG-5'
b
Port
Pillars
Nanochannels
c
DNA Mass (A.U.)
0
5
10
15
20
25
30
35
40
0
50
100
150
200
250
DNA Molecule Length (Kb)
d

**Figure 5.3 Nano-mapping of mixtures of 95 BACs from the PGF and COX libraries.** (**a**) Image of a single field of view (FOV of 73x73 microns) containing a mixture of nick labeled DNA molecules in the nanoarray. This FOV is part of 108 FOVs shown in the bottom part of the panel (outlined in green). Each FOV can accommodate up to 250 kb of a DNA molecule from top to bottom. The images of 4 FOVs are stitched together so that longer molecules (up to 1 Mb) in a single channel can be analyzed whole. In all, there are 27 sets of 4 vertical FOVs per array scan. (**b**) The distribution of the DNA molecules imaged on the nanoarray by length. As shown in the plot of number of DNA molecules at each molecular length, the majority of the molecules are 100-170 kb in length as expected from the BAC-clone sizes. (**c**) After clustering of DNA molecules based on nick-labeling patterns, consensus maps with overlapping patterns are assembled into contiguous sequence motif maps. In this example, 3 overlapping consensus maps (each ~150 kb long) are assembled into a 300 kb map.

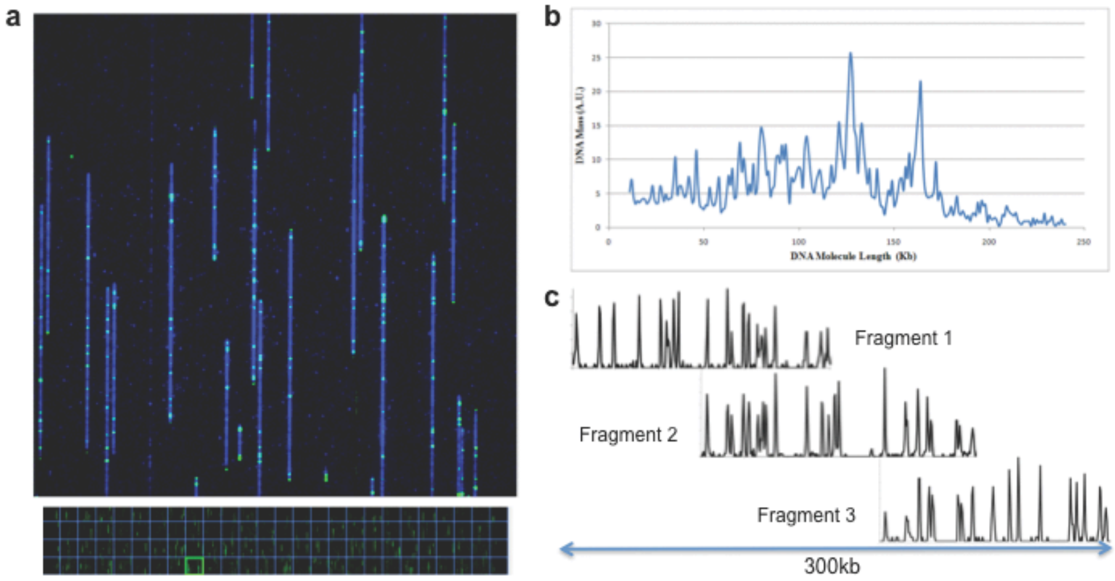


**Figure 5.4 Sequence motif map of the MHC region.**
The *in silico* sequence motif map for PGF reference is represented by a black line with the Nt.BspQI sites marked with black dots and the map of the same region produced by nano-mapping is represented by a blue line with blue dots. Where there are motif variations between COX and PGF, the COX motif is represented with red lines and red dots. Asterisks mark the gaps in the Nt.BspQI map produced by nano-mapping. Gene locations and the location of the variable RCCX module are noted. Additional loci of special interest are marked with boxes and are discussed in detail in the text. Green line indicates the region drawn in Figure 5.6.

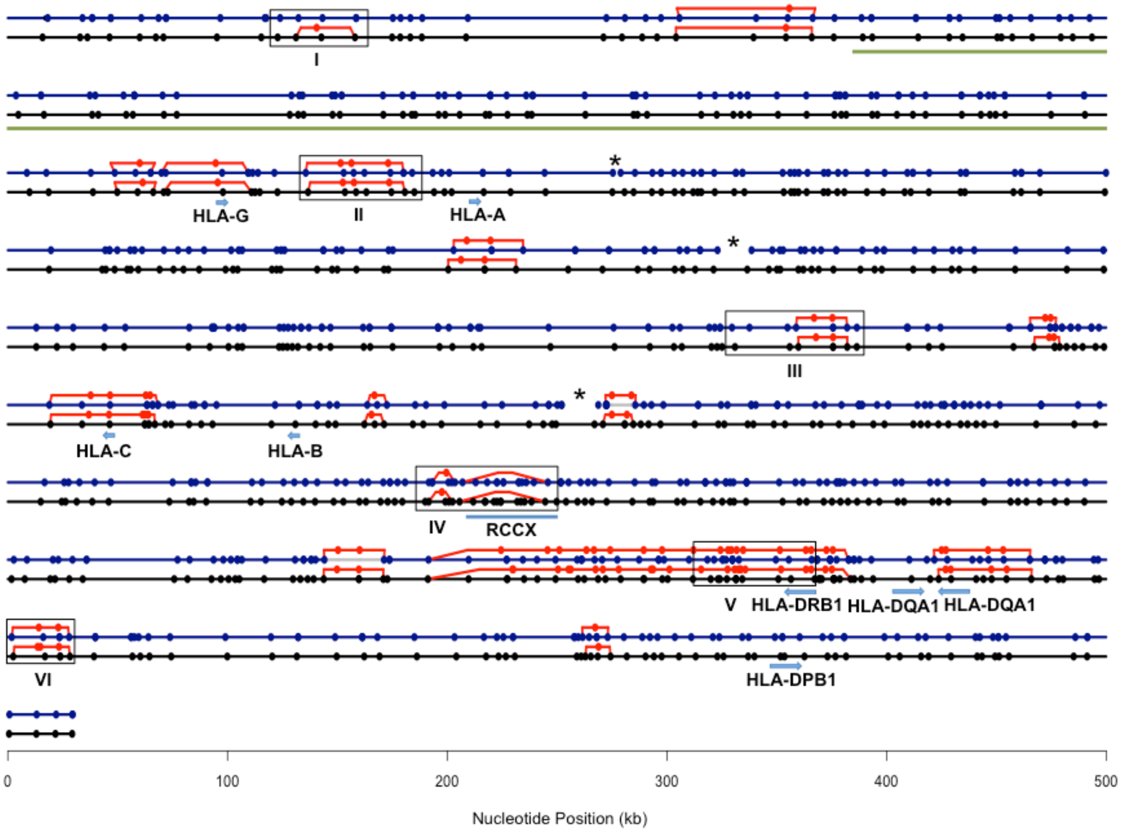


**Figure 5.5 Discrepancies between the reference Nt.BspQI map and that produced by nano-mapping.**
(**a**) The reference Nt.BspQI maps of the region (Box I in Fig. 4) indicate that the COX genome (grey line with red dots) has a 4 kb deletion as compared to the PGF genome (grey line with blue dots), with 7 kb vs 11 kb fragment between two neighboring sites. The map of the same region produced by nano-mapping from both libraries (histogram plot in black) shows the same haplotype for both COX and PGF genomes, with an 11 kb fragment between the corresponding 2 sites. (**b**) An Nt.BspQI site identified in the region marked as Box III in Fig. 4 (arrow) is found in the PGF genome (blue histogram plot) by nano-mapping, splitting the 24 kb fragment in the reference map (black line) into 7 kb and 17 kb fragments. The COX reference map (red line) and the COX map produced by nano-mapping (red histogram plot) are also displayed to show that the COX genome has the 24 kb fragment and a haplotype variation in the adjacent region.

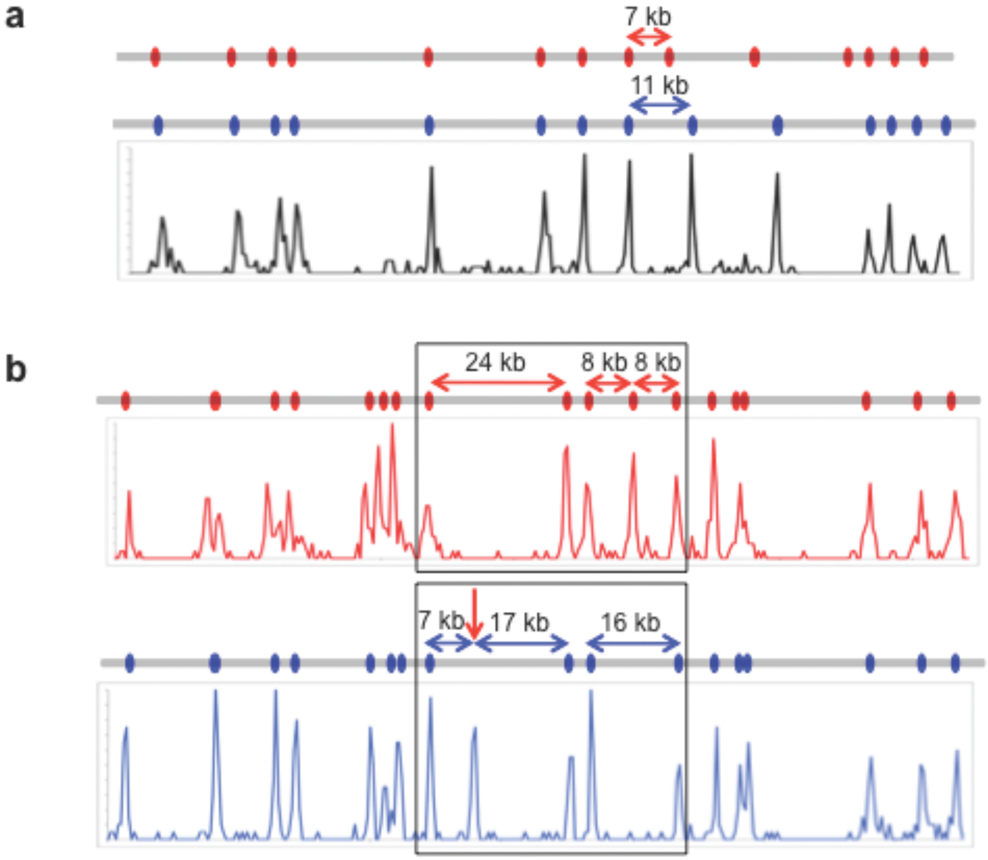


**Figure 5.6** ***De novo*** **sequence assembly of the MHC region from the PGF and COX genomes.**
DNA of 95 BACs from the PGF and COX libraries was sequenced and the sequence reads were assembled into contigs (arrows) with dots marking the Nt.BspQI sites. The contigs were aligned to the Nt.BspQI map produced by nano-mapping, providing information on the relationship and orientation of contigs together with the location and size of each gap between contigs. Shown are *in silico* sequence motif maps of the contigs and of the reference sequence to emphasize accuracy of the assembled contigs.

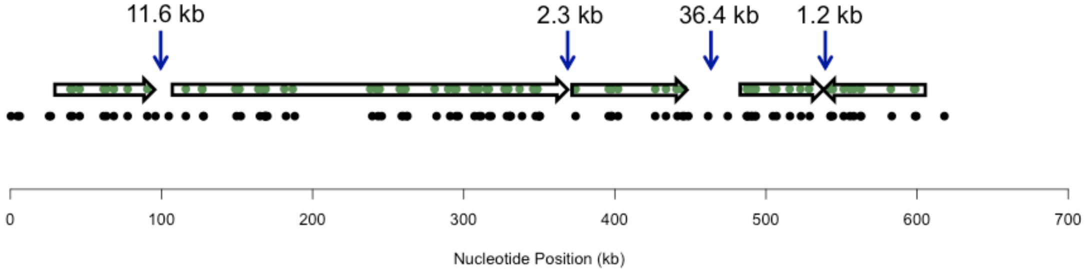
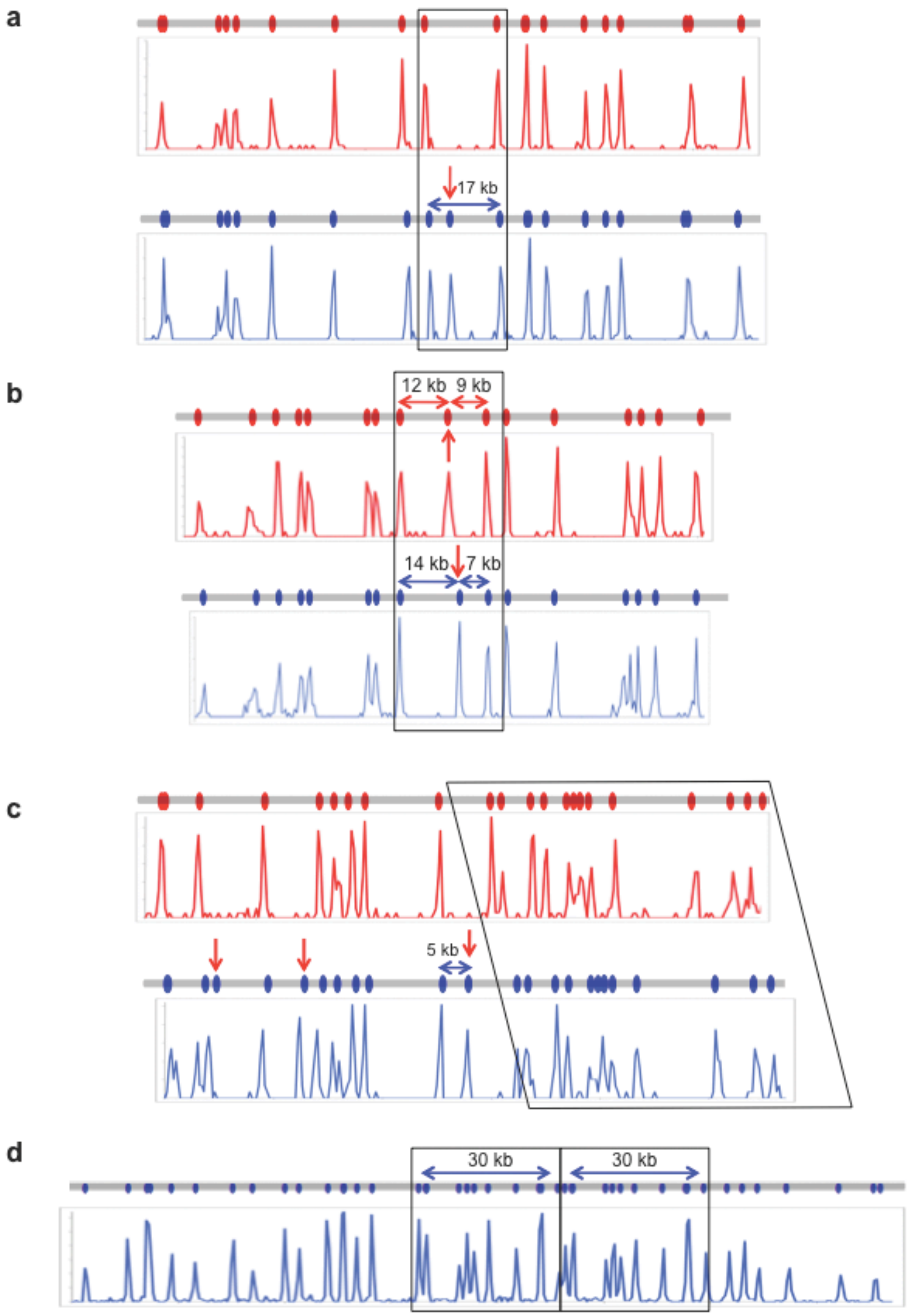


**Figure 5.7 Haplotype resolution and structural variation detected by nano-mapping.** (**a**) Single-site variation due to the creation or destruction of an Nt.BspQI site is easily identified by nano-mapping. The region marked as Box II in Fig. 4 shows that the PGF genome (blue line) contains an extra Nt.BspQI site not found in the COX genome (red line) with the maps generated by nano-mapping (blue and red histogram plots) showing the expected pattern. (**b**) Shifting of a site relative to others in two haplotypes may be due to a double mutation or an inversion event. In the region marked as Box VI in Fig. 4, the 21 kb region is split into 12/9 kb fragments in the COX genome (red line and red histogram plots) but 14/7 kb fragments in the PGF genome (blue line and blue histogram plot). (**c**) Insertions can be identified and localized by nano-mapping for haplotyping resolution. In the region marked as Box V in Fig. 4, the PGF genome has a 5 kb insertion that also includes an Nt.BspQI site (blue line, blue histogram plot) when compared to the COX genome (red line, red histogram plot). (**d**) A 30 kb duplication at the RCCX locus (Box IV in Fig. 4) is easily identified and localized in both the reference map (black line) and that produced by nano-mapping (blue histogram plot).


a
17 kb
b
12 kb
9 kb
14 kb
7 kb
c
5 kb
d
30 kb
30 kb

**Publishing Agreement**

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

ErnestLam. June 15, 2012

**Ernest Lam** **Date**