GENOTYPE-PHENOTYPE ASSOCIATION USING HIGH THROUGHPUT SEQUENCING DATA

by

Zev Nachman Kronenberg

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

August 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Zev Nachman Kronenberg**

has been approved by the following supervisory committee members:

Mark Yandell , Chair **April 23, 2015**
Date Approved

Mark Metzstein , Member **April 23, 2015**
Date Approved

Nels Elde , Member **April 23, 2015**
Date Approved

Karen Eilbeck , Member **April 23, 2015**
Date Approved

Michael Shapiro , Member **April 23, 2015**
Date Approved

and by **Lynn Jorde** , Chair/Dean of
the
Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Genotype Phenotype Association (GPA) is a means to identify candidate genes and genetic variants that may contribute to phenotypic variation. Technological advances in DNA sequencing continue to improve the efficiency and accuracy of GPA. Currently, High Throughput Sequencing (HTS) is the preferred method for GPA as it is fast and economical. HTS allows for population-level characterization of genetic variation, required for GPA studies.

Despite the potential power of using HTS in GPA studies, there are technical hurdles that must be overcome. For instance, the excessive error rate in HTS data and the sheer size of population-level data can hinder GPA studies.

To overcome these challenges, I have written two software programs for the purpose of HTS GPA. The first toolkit, GPAT++, is designed to detect GPA using small genetic variants. Unlike pervious software, GPAT++'s association test models the inherent errors in HTS, preventing many spurious GPA. The second toolkit, Whole Genome Alignment Metrics (WHAM), was designed for GPA using large genetic variants (structural variants). By integrating both structural variant identification and association testing, WHAM can identify shared structural variants associated with a phenotype. Both GPAT++ and WHAM have been successfully applied to real-world GPA studies.

TABLE OF CONTENTS

Chapters

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge the patience of both my past and current advisor, Holly Wichman and Mark Yandell, as I am a stubborn student. I am also grateful to the Genetics Training Grant community, the Department of Human Genetics, and my committee members for outstanding mentorship. I would also like to thank my colleagues, Matt Settles, Sam Hunter, Edward Osborne, and Ryan Abo, for feedback on my projects.

The encouragement I have received from my family has been immeasurable. My wife, Ellen Kronenberg, has made many sacrifices, so that I can pursue a scientific career.

CHAPTER 1

INTRODUCTION

1.1 Identifying the genetic basis of phenotypic variation

Phenotypic variation is any measureable difference between two individuals within a population. An astounding amount of phenotypic variation can exist within a single species. For example, there are over 300 different breeds of domestic rock pigeons with varying size, weight, pigmentation, and plumage patterns, all derived from a single ancestral phenotype (Figure 1.1). There are two general classifications for phenotype variation, continuous and discrete traits. Continuous traits, like weight or height (complex traits), have continuous distributions, meaning they can take on any value within a range, while discrete traits (simple traits) only fall into a few categories. An example of a discrete trait is the head crest, which is present in some pigeon breeds, but not others (Figure 1.1)(Shapiro et al., 2013a). Phenotypic variation is not limited to visible differences, but can also include molecular phenotypes. For example, within the human population, genetic variation in the bitter taste receptor gene, TAS2R, confers the ability to taste PTC (Kim et al., 2003). Another example of a molecular phenotype in humans is the blood group determination system (Owen, 2000). Phenotypic variation can be passed from parents to offspring. The amount of phenotypic variation explained by genetics is called heritability (Visscher et al., 2008). Many simple traits like hair or eye color have high heritability, while other traits, like weight and height, have low heritability.

The quest to understand the genetic basis of phenotypic variation has been ongoing for over a century. Open-ended questions in the field are: How do environmental factors affect the phenotypic outcome of genetic variation? How do genetic variants interact with each other (epistasis)? Which classes of variants are beneficial, conferring adaptive potential, and which classes are deleterious or disease causing? Early geneticists began working on these questions by studying the heritability of traits in crossbreeding experiments. For example, Gregor Mendel, an Augustine friar, established the rules of heritability by observing the outcome of crossing pea plants with different phenotypes.

Unlike Mendel's experiments, later work showed that two phenotypes can be transmitted together and thus suggests they are linked genetically. Alfred Sturtevant, a student of Thomas Hunt Morgan, pioneered linkage/genetic mapping in *Drosophila* by crossbreeding flies with different phenotypes (Griffiths et al., 1999). Linkage analyses track the co-occurrence of phenotypes over reproductive generations. During sexual reproduction, in many organisms, sections of two homologous chromosomes cross over, redistributing the genetic variants responsible phenotypic variation. The frequency at which two phenotypes co-occur in the progeny of a genetic cross is thus inversely proportional to the physical distance of the causal variants on a chromosome. Genetic variants within genes that are in close physical proximity tend to segregate together, while recombination breaks the linkage between distant genes. Carefully tracking the co-occurrence of many phenotypes over many crosses allows for a genetic map to be built without any knowledge of physical position of the genes (physical map).

The physical map describes the location of genetic attributes on a chromosome; landmarks include genes and restriction enzyme cut sites. Linkage maps and physical maps are positively correlated. However, the relationship is not linear. Along a chromosome, there are regions of low recombination where the genetic distance increases at a slower rate than the physical distance.

Recent advances in automated DNA sequencing provided the means to characterize the physical map of the human genome at base pair level (McPherson and Others, 2001). The Human Genome Project, completed in 2003, provided the full sequence of the human genome and annotated ~20,500 genes (Hattori, 2005; Lander et al., 2001). This breakthrough allowed for human linkage analyses to be overlaid upon the physical map, accelerating the rate at which phenotypes and genes could be linked. The Human Genome Project also allowed for standing genetic variation within the human population to be mapped to a physical location. The International HapMap Project used automated sequencing to survey the genetic diversity (bi-allelic Single Nucleotide Polymorphisms [SNPs]) across unrelated individuals of Yoruban, Chinese, European, and Japanese ancestry (The International HapMap Consortium, 2005). By capturing a wide swath of haplotype diversity in the human population, the International HapMap Project provided a set of genetic markers linked to the physical map, a resource that has been widely used for genome-wide association studies in the years since.

1.2 Genome-wide association studies (GWASs) in the SNP chip era

Genome-wide association studies were developed shortly after the completion of the Human Genome Project and the International HapMap Project. Array-based genotyping (SNP chips) used evenly spaced HapMap SNPs (at least one SNP every 5Kbps) to create genotype markers across the genome (The International HapMap Consortium, 2005). Early iterations of these arrays only contained tens of thousands of markers, but modern chips represent over a million SNP loci (LaFramboise, 2009). SNP chips make genotyping hundreds of humans financially feasible (Spencer et al., 2009). Genotypes called from SNP arrays are highly accurate. For example, the genotypes called from Affymetrix arrays have concordance rates above 0.99 for HapMap data (Rabbee and Speed, 2006). The low costs of SNP chips and high-quality data obtained have allowed geneticists to carry out large GWASs for a variety of medically relevant phenotypes (Altshuler, 2009). GWASs work by measuring the association between phenotypes and the genotypes at common SNP loci (Hancock and Scott, 2012). There are two subclasses of GWAS, one that is designed for discrete phenotypic variation and the other that is designed for continuous phenotypic variation, or quantitative traits. GWAS for simple traits requires two cohorts, an affected group (called cases or target) and a healthy group (called controls or background). The allele frequencies of the cases and controls are compared at each marker to test for an association. GWAS for continuous traits does not require two cohorts, but instead measures individuals' phenotype(s) and groups these phenotypic measures by genotypes. If the grouped phenotypic values differ between genotypes then the genotype may affect the phenotypic variance. To date, GWAS has been successful in identifying regions of the genome that are associated with human maladies such as Crohn's disease, heart disorders, and obesity, to name a few (Welter et al., 2014).

There are many statistical and biological considerations in GWAS design. It is important to consider the factors that can confound true and false associations. In simple trait association studies, population stratification between cases and controls can cause false positives (Tian et al., 2008). For example, if the affected cohort is comprised of Caucasians and the background is made up of people with African ancestry, there will be a high degree of allelic differentiation between the two groups that has nothing to do with the phenotype in question. Ideally, the cases and controls should be genetically indistinguishable, except for the locus/loci causing the phenotypic difference(s). In other words, they should be a single homogeneous population. There are several common methods used to test for population stratification. Principle component analyses (PCA) can unveil hidden stratification within the cases and controls (Patterson et al., 2006). Cryptic relatedness between individuals in a study can also generate false positives as related individuals share large haplotypes that inflate association (Astle and Balding, 2009). There are regional methods that can identify shared genomic segments (SGS) between individuals (Knight et al., 2012). The statistical models in GWAS have also been modified to directly account for population stratification and cryptic relatedness. The linear models in GWAS can be adjusted for co-variants including age, sex, weight, and other common vital signs. Another approach to GWAS is to look for regional ancestry differences between cases and controls, termed admixture mapping (Winkler et al., 2010). If care is taken during GWAS design, it can be a very powerful tool for identifying candidate genomic intervals associated with a phenotype of interest.

SNP chip-based GWAS has several shortcomings. The most inherent drawback of SNP GWAS is that it only identifies the causative loci with neighboring marker SNPs through linkage disequilibrium. This is an indirect association as a GWAS does not identify the causative variant, but localizes the association, through the principle of linkage disequilibrium, to a genomic neighborhood. The second drawback is that GWAS is dependent on recombination rates (Visscher et al., 2012). The resolution of genetic mapping depends on the number of recombination events across the region. If a hypothetical causative variant is within a low recombination interval, it will take many more samples to recover enough recombination events to fine map the variant. Ascertainment bias is another inherent and serious drawback of SNP chip GWAS (Albrechtsen et al., 2010). The genetic variation present in some human populations is not well represented on the SNP chips. As a result, the allele frequency spectra are skewed toward common alleles. Severe ascertainment bias can reduce the power of GWAS when informative markers are not included on the chip. Lastly, rare variants are poorly represented on SNP chips (Zeggini et al., 2005). This poses a problem because these variants often have low linkage disequilibrium with the SNPs present on the chip, meaning very large sample sizes are needed to generate an association signal. Each of these drawbacks must be accounted for when designing a GWAS experiment.

1.3 Genotype Phenotype Association with HTS

In the mid 2000s, high throughput sequencing (HTS) methods became widely available (Mardis, 2008). Technical advances in molecular biology, microfluidics, and optical imaging paved the way for HTS technologies. Companies like 454 Life Sciences (now Roche), Illumina, PacBio, and Oxford Nanopore all offer unique variations of HTS. Over the past decade, the costs of HTS have plummeted and it is projected that sequencing a whole human genome will soon cost less than $1,000. Because of low HTS costs, studies designed to link genetic variants to phenotypes often use HTS over SNP arrays.

There are several advantages of using HTS for GWAS over SNP chips. The most obvious difference between these technologies is that HTS can identify the causative genetic variants underlying a phenotype or disease. Directly linking genetic variants and phenotypes is called Genotype Phenotype Association (GPA). Again, SNP chips only provide a set of markers; therefore, they do not directly identify causative variants. The second difference is that, unlike SNP chips, HTS does not have an ascertainment bias. On average, roughly three million variants are identified when a new human genome is sequenced and compared to a reference sequence. This allows for the unbiased identification of both rare and novel variants, including *de novo* mutations. The unbiased nature of whole genome sequencing also allows for structural variants ranging from 1bp up to several megabases to be identified (Abel and Duncavage, 2013). Before HTS, array comparative genomic hybridization and tiling-based arrays were used to identify structural variation. These methods are expensive and have poor resolution for smaller structural variants (Coe et al., 2007). For these reasons, HTS has become the preferred technology for GPA. The genetic basis of many phenotypes and diseases has been established using HTS (Bahassi and Stambrook, 2014).

HTS GPA can be performed using both whole genome and exome sequencing data. Exome sequencing is accomplished by capturing the ~1% of the genome that is protein coding using hybridization techniques (Biesecker and Green, 2014). Exome sequencing has two advantages over whole genome sequencing (WGS). First, exome sequencing analyses are constrained to the better-understood regions of the genome. The functional implications of single nucleotide variants and structural variant within coding sequences are easier to predict than noncoding variants. The second advantage is financially motivated; exome sequencing is ~1/6th the cost of whole genome sequencing. Increased sampling further increases the power to associate variants with phenotypes. In contrast to exome sequencing, WGS allows for the analysis of noncoding variants. There is growing awareness that noncoding variants are functionally relevant in evolution and disease and should not be ignored (Maurano et al., 2012; Wray, 2007).

The role of rare variants in common diseases can be quantified using both WGS and exome sequencing. Burden tests have been developed over the last decade to detect associations between phenotypes and rare variants contained with a gene. Diseases or phenotypes that are causes by many unique alleles in the same gene cannot be detected by single marker tests, as there is little to no genetic sharing. Burden tests summarize the load of predicted deleterious variants within a genomic feature. In most cases, a feature is defined as the coding region of a gene. Burden tests only require one statistical test per gene, whereas single marker tests require a statistical test for every genetic variant. Burden tests are not dependent on allelic sharing to detect an association; however, shared alleles increase the power of some burden tests. SKAT, VAAST, and C-alpha are all programs that can conduct a burden test (Neale et al., 2011; Wu et al., 2011; Yandell et al., 2011). Burden tests can also be extended to *de novo* genetic variants for small pedigrees (trios and quartets). Under a *de novo* model, all variants shared between a proband and their family members can be excluded. In the case of large pedigree analyses, pVAAST can incorporate signals of linkage into its burden test to gain statistical power (Hu et al., 2014). Burden tests have been widely adopted in genotype-phenotype association studies.

Another approach to identifying genetic basis of a phenotype is variant prioritization. Variant prioritization tools categorize genetic variants as damaging, benign, or somewhere in-between. Variant prioritization tools integrate amino acid substitution frequencies, allele frequency, and phylogenetic conservation to annotate the affect of mutations and rank damaging alleles. Sift, Annovar, CADD, and PolyPhen are just a few of the popular tools in use today (Adzhubei et al., 2010; Kircher et al., 2014; Ng, 2003; Wang et al., 2010). These tools generally produce allele-specific scores rather than a p-value from a statistical test. Researchers then scour lists of ranked variants to find candidates for the phenotypes they are studying. By applying *a priori* knowledge, in the form of a gene candidate list, the number of potential candidates can further be reduced. Variant prioritization remains a popular approach as it is intuitive and provides researchers the ability to sort candidates based on their biological knowledge.

HTS has opened the world of GPA and genomics to many model and nonmodel and species. Before HTS, creating a physical map and designing a SNP chip for GPA in a new species was prohibitively expensive. It is now common for hundreds of *de novo* genomes to be published each year, accompanied by population level resequencing projects (Ekblom and Galindo, 2011; Ellegren et al., 2012). Alleles controlling phenotypic variation in rice, chickens, cows, and dogs have all been identified using HTS GPA (Freedman et al., 2014; Huang et al., 2010; Jansen et al., 2013; Rubin et al., 2010). GPA studies in domesticated species have been wildly successful because of strong artificial selection. During domestication, humans have often selected for one or a few alleles per phenotype of interest. It is easier to identify a GPA when there are only a few alleles controlling the trait of interest. GPA studies in domestic systems are being used to further basic biological research and increase the effectiveness of animal husbandry.

1.4 The challenges of Genotype Phenotype Association (GPA) using HTS

There are several unique sources of error that can affect HTS GPA studies. These errors can be divided into four general categories: sequencing errors, mapping errors, genotyping errors, and pipeline bias. If GPA studies do not control for these errors, they can suffer high false discovery rates or fail to identify causative variant(s). GPA methodology must keep evolving to cope with new types of HTS data and the associated errors. In this section, I will discuss the four types of errors and ways to prevent error propagation.

Sequencing error, also known as base calling error, occurs when a nucleotide is misread either by machine or by human. Automated HTS base calling has an error rate ten times higher than Sanger sequencing (Kircher and Kelso, 2010). For example, in Illumina data, base calling errors occur at a rate of 0.26% to 0.8% depending on the platform (Quail et al., 2012). HTS error rates are not constant across a read. Errors are more common at the start and end of a read due to the sequencing chemistry (Minoche et al., 2011). Long read technologies, such as pacBio SMRT sequencing, have base calling error rates around 10-14%, which is two orders of magnitude greater than Illumina sequencing technologies (Roberts et al., 2013). There are also sequence-specific errors; repeats of ‘GCC’ and ‘GGC’ cause Illumina sequencing errors (Nakamura et al., 2011; Quail et al., 2012). HTS errors occur randomly, with the exception of sequence-specific errors. Therefore, sequencing a genome to a higher depth can mitigate the affects of base calling errors. An alterative strategy for dealing with base calling errors is to trim the 5’ and 3’ of reads, removing many erroneous nucleotides (Schmieder and Edwards, 2011).

Mapping errors are a serious and pervasive form of error that affects all GPA analyses. HTS reads derived from complex or highly repetitive regions of the genome cannot be uniquely mapped and are randomly assigned (Li, 2014)(Figure 1.2A-B). Reads derived from high copy number sequences, such as transposable elements including LINE-1 and ALU elements, are often ambiguously mapped. Reads with real genetic variation that are incorrectly mapped introduce false variant calls at the positions where they were placed. Reads sampled from real genetic variants, such as insertions, cannot be mapped because the insertion sequence is not represented in the reference genome. Structural variants, including deletions and inversions (Figure 1.4.1C-D), can induce false variant calls when their breakpoints are not correctly identified. There are three ways to reduce the number of mapping errors. First, mate-pair sequencing allows mapping software to find genomic positions where both mates map. If one fragment is present in a repetitive sequence, its mate can rescue it to the correct location. Second, as HTS reads get longer, the ability to map the reads will improve. Third, efforts to identify and re-assemble erroneous regions in reference genomes will reduce mapping errors.

Genotyping errors are a direct consequence of sequencing and mapping errors. Variant callers rely on the mapping qualities and base qualities provided by upstream tools (Nielsen et al., 2011). Modern variant callers integrate many sources of information into their probabilistic models (Li, 2011; McKenna et al., 2010). However, even with these models, errors are common. Downstream filtering is a common way to reduce the number of false positive variant calls. Common filters are depth, genotype likelihoods, and frequency of the variant. Population genetic metrics such as Hardy Weinberg Equilibrium have also been widely used. Genotyping algorithms can also leverage pedigree- or population-level allele frequencies to improve the quality of calls. For example, in the case of a trio pedigree, the rules of inheritance can be applied to reduce false variant calls, except in the case of *de novo* mutations. Insuring that only high-quality genotype calls are used for GPA will reduce the amount of spurious associations.

The last source of errors is pipeline/logistical errors. These errors include everything from study design, to sample preparation, to the software tools used for the analyses (O’Rawe et al., 2013; Robasky et al., 2014). When designing a GPA, it is imperative that enough samples are gathered for the study to tolerate human errors such as mislabeling. Other sources of logistical error include mechanical malfunctions during sequencing and hard drive failures. Pipeline biases can occur when multiple samples are processed with different alignment software, or variant callers. This last source of error is common when small research groups compare their genomes to large consortiums like One Thousand

Genomes Project or the Exome Aggregation Consortium (ExAC)(Abecasis et al., 2012). Significant effort should be expended to ensure that both the cases and controls of a GPA are processed in the same way.

1.5 Summary of chapters

In this thesis, I present several examples applying HTS for Genotype Phenotype Association. We have developed two software suites for this purpose. The first tool, GPAT++, discussed in Chapter 2, is used for Genotype Phenotype Association for bi-allelic SNPs and indels. GPAT++ was extensively applied to biological data in Chapters 3 and 4. In Chapter 5, I introduce WHAM, a tool designed for structural variant GPA. Unlike GPAT++, WHAM does not rely on variant calls from other tools, but rather directly calls structural variants and conducts association testing from binary alignment files. The development and application of GPAT++ and WHAM represent the majority of my Ph.D. efforts.

1.5.1 GPAT++

The application of population genomics to nonmodel organisms is greatly facilitated by the low cost of next-generation sequencing (NGS), including methods that seek to overcome some of the problems discussed above. Barriers, however, exist for using NGS data for population level analyses. Traditional population genetic metrics, such as F_{st}, are not robust to the genotyping errors inherent in noisy NGS data. Additionally, many older software tools were never designed to handle the volume of data produced by NGS pipelines. To overcome these limitations, we have developed a flexible software library designed specifically for large and potentially noisy NGS datasets. The Genotype Phenotype Association Toolkit (GPAT++) implements both traditional and novel population genetic methods in a single user-friendly framework. GPAT++ consists of a suite of command-line tools and scripts that enable rapid genotype-phenotype association with visualization methods to aid analyses. The methods implemented in GPAT++ have been applied to several projects presented in subsequent chapters of this thesis. We have used GPAT++ on a variety of non-model systems including, but not limited to, pigeon, pine fungus, poxvirus, and *Tetrahymena*.

At the core of the GPAT++ suite is a single marker genotype-phenotype association test called pFst. The GPA tool, pFst, is comprised of two separate likelihood ratio tests that work on diploid organisms (genotypic data) or pooled sequencing data (allele frequency data). Unlike previous GPA statistical tests, pFst integrates genotype likelihoods for nonpooled data. Modeling the genotyping errors, via the genotype likelihoods, helps to reduce the numbers of false positives. pFst can process tens of millions of variant sites in a variant call file (VCF) within a few hours. For these reasons, pFst is an excellent assay for initial association testing.

In this chapter, I will describe, in detail, the statistical test pFst and the implementation. The code is publically available on Github and a detailed wiki provides the necessary user documentation to run the code.

1.5.2 Genomic diversity and the evolution of the head crest in the rock pigeon

The phenotypic variation present in domestic rock pigeons is astounding. In 2013, we published the rock pigeon genome along with the genome re-sequencing of more than 30 phenotypically diverse breeds (Shapiro et al., 2013b). This single dataset has been leveraged to map the genetic basis of head-crest and several other traits. At the time, it was one of the earliest examples of using whole-genome resequencing to map the genetic basis of an avian trait. This dataset spurred us to write the GPAT++ suite, as many of the tools that were available were woefully ill-suited for 40 whole genomes.

In the manuscript, we describe mapping the allele that caused the recessive headcrest trait using two complementary approaches. Traditional metrics of natural selection including F_{st} and XP-EHH were used to identify a candidate region (scaffold 612) based on allele frequency differences and haplotype structure around the headcrest locus. The Variant Annotation And Search Tool (VAAST) is a disease gene-finding tool that was coopted for GPA. VAAST's ability to integrate differences in allele frequencies between cases and controls, phylogenetic conservation, models of amino-acid substitutions, and models of inheritance makes VAAST a powerful tool for GPA. VAAST identified a single amino acid substitution in the *EphB2* gene as the best candidate for the headcrest phenotype genome-wide. Additional Sanger-based genotyping bolstered the perfect association of the allele in *Ephb2* and the head-crest phenotype.

In addition to mapping headcrest, we used the 40 genomes to model the evolutionary history of the domestic pigeon, including the effective population size and relationships between the breeds. We found genetic introgression between birds that supported ancient human trade routes between different ethnic groups. Phylogenetic analyses of the birds revealed that artificial selection from breeders has affected the genetic diversity across all domestic pigeons. We found within-breed diversity was low while between-breed diversity is quite high. This supported the idea that pigeon breeders have maintained the "purity" of breeds over time.

1.5.3 Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon

Pigmentation plays an important role in mate choice, predator avoidance, and mimicry in many natural populations. In domestic species like the pigeon, breeders' preference often determines if a bird's plumage coloration is favorable. Domestic pigeon breeds have a variety of pigmentation and patterning. Black, brown, yellow, red, and blue are all colors that are observed across the breeds. These colors often appear to possess epistatic interactions with one another. In lay terms, when two different color alleles are present in the same bird, the phenotype will be discrete rather than a blending of colors. Chapter 4 presents the paper "*Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon*" (Domyan et al., 2014). In this paper, we discovered three genes, *Tryp1*, *Sox10*, and *Slc45a2*, as well as several alleles within these genes that interact to generate six unique plumage colors. Some of the alleles were genic (*Tryp1* and *Slc45a2*), while two alleles, *e1* and *e2*, were deletions of a melanocyte specific enhancer upstream of *Sox10*. We used pFst to identify the association of recessive red with the region upstream of *Sox10* and through careful analysis, the noncoding *e1* and *e2* alleles were discovered. VAAST was applied to identify the coding variants within *Slc45a2* and *Tyrp1*. The two tools, VAAST and pFst, are complementary as VAAST has excellent power in coding regions, while pFst excels at identifying noncoding variants associated within phenotypes. This paper demonstrates that pigeon genetics and genomics can be used to unravel the interactions between genes and alleles.

1.5.4 DisAp-dependent striated fiber elongation is required to organize ciliary arrays

Tetrahymena thermophila is a free-living ciliate that serves as a good model system for dissecting one of the most complex eukaryotic molecular motors, the cilium. *Tetrahymena* are covered from tip to tail with cilia, which are required for cell motility. Genetic screens have been carried out in *Tetrahymena* to identify and characterize mutations that cause motility defects. These defects are of great interest to molecular biologists as they mimic primary ciliary dyskinesia, a serious human disease where cilia fail to clear mucus from the respiratory tract (Noone et al., 2004).

Genetic screens in the late 1970s discovered a *Tetrahymena* mutant with basal body defects resulting in the disorganization of cilary arrays. The mutation, that caused the phenotype, was recessive and called *disA-1*. I used GPAT++ to map the disA-1 locus, which allowed our colleagues to genetically and molecularly characterize the role of the novel protein DisAp (Galati et al., 2014).

1.5.5 WHAM: Identifying structural variants of biological consequence

Genotype phenotype association (GPA) testing using high throughput sequencing data has traditionally excluded structural variants (SV) due to inaccuracies during SV detection. High rates of false positives and negatives are common among the state-of-the-art tools. Most SV detection tools error on the side of specificity at the expense of sensitivity, meaning they often fail to call true structural variants. This aspect of SV detection is highly detrimental for GPA as SV alleles associated with a phenotype can be missed. To incorporate SV calls into genotype phenotype association, there are three challenges that must be addressed. First, the breakpoints of SVs must be reliably identified across multiple individuals. Second, the breakpoints must be genotyped correctly. Lastly, genotyping and breakpoint identification errors must not be systematic between the cases and controls. Error must be random and even across the case and control cohorts.

To address these issues, we have created an association-testing tool for structural variants entitled: **Wh**ole **G**enome **A**lignment **M**etrics, or WHAM for short. WHAM is a population-level structural variant caller that also conducts genotype-phenotype association testing. WHAM is highly sensitive with respect to SV identification. WHAM provides high positional breakpoint accuracy and reliable genotype calls. In this chapter, I present WHAM benchmarked on both simulated and biological datasets. As a proof of principle, WHAM was applied to re-discover the *e1* allele, a deletion that is responsible for the recessive red phenotype in pigeons (see Chapter 3).

1.6 References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M. a, Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G. a (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abel, H.J., and Duncavage, E.J. (2013). Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. *206*, 432–440.

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. *27*, 2534–2547.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881–888.

Astle, W., and Balding, D.J. (2009). Population structure and cryptic relatedness in genetic association studies. Stat. Sci. *24*, 451–471.

Bahassi, E.M., and Stambrook, P.J. (2014). Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis *29*, 303–310.

Biesecker, L.G., and Green, R.C. (2014). Diagnostic clinical genome and exome sequencing. N. Engl. J. Med. *370*, 2418–2425.

Coe, B.P., Ylstra, B., Carvalho, B., Meijer, G. a, Macaulay, C., and Lam, W.L. (2007). Resolving the resolution of array CGH. Genomics *89*, 647–653.

Domyan, E.T., Guernsey, M.W., Kronenberg, Z., Krishnan, S., Boissy, R.E., Vickrey, A.I., Rodgers, C., Cassidy, P., Leachman, S. a, Fondon, J.W., et al. (2014). Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr. Biol. *24*, 459–464.

Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb). *107*, 1–15.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., et al. (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature *491*, 1–5.

Freedman, A.H., Gronau, I., Schweizer, R.M., Ortega-Del Vecchyo, D., Han, E., Silva, P.M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., et al. (2014). Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. *10*, e1004016.

Galati, D.F., Bonney, S., Kronenberg, Z., Clarissa, C., Yandell, M., Elde, N.C., Jerka-Dziadosz, M., Giddings, T.H., Frankel, J., and Pearson, C.G. (2014). DisAp-dependent striated fiber elongation is required to organize ciliary arrays. J. Cell Biol. *207*, 705–715.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., and Lewontin, R.C. (1999). Modern genetic analysis (WH Freeman New York).

Hancock, D.B., and Scott, W.K. (2012). Population-based case-control association studies. In Current Protocols in Human Genetics, *74*, 1.17.1–1.17.20.

Hattori, M. (2005). Finishing the euchromatic sequence of the human genome. Tanpakushitsu Kakusan Koso. *50*, 162–168.

Hu, H., Roach, J.C., Coon, H., Guthery, S.L., Voelkerding, K. V, Margraf, R.L., Durtschi, J.D., Tavtigian, S. V, Shankaracharya, Wu, W., et al. (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat. Biotechnol. *32*, 663–669.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. *42*, 961–967.

Jansen, S., Aigner, B., Pausch, H., Wysocki, M., Eck, S., Benet-Pagès, A., Graf, E., Wieland, T., Strom, T.M., Meitinger, T., et al. (2013). Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics *14*, 446.

Kim, U., Jorgenson, E., Coon, H., Leppert, M., Risch, N., and Drayna, D. (2003). Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. Science *299*, 1221–1225.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. Bioessays *32*, 524–536.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Knight, S., Abo, R.P., Abel, H.J., Neklason, D.W., Tuohy, T.M., Burt, R.W., Thomas, A., and Camp, N.J. (2012). Shared genomic segment analysis: the power to find rare disease variants. Ann. Hum. Genet. *76*, 500–509.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. *37*, 4181–4193.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843–2851.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. *9*, 387–402.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science (80-.). *337*, 1190–1195.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McPherson, J.D., and Others (2001). A physical map of the human genome. Nature *409*, 934.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. *12*, R112.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. *39*, e90.

Neale, B.M., Rivas, M. a, Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.

Ng, P.C. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Noone, P.G., Leigh, M.W., Sannuti, A., Minnix, S.L., Carson, J.L., Hazucha, M., Zariwala, M. a, and Knowles, M.R. (2004). Primary ciliary dyskinesia: diagnostic and phenotypic features. Am. J. Respir. Crit. Care Med. *169*, 459–467.

O’Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. *5*, 28.

Owen, R. (2000). Karl Landsteiner and the first human marker locus. Genetics *155*, 995–998.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Quail, M. a, Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Rabbee, N., and Speed, T.P. (2006). A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics *22*, 7–12.

Robasky, K., Lewis, N.E., and Church, G.M. (2014). The role of replicates for error mitigation in next-generation sequencing. Nat. Rev. Genet. *15*, 56–62.

Roberts, R., Carneiro, M., and Schatz, M. (2013). The advantages of SMRT sequencing. Genome Biol *14*, 405.

Rubin, C.-J., Zody, M.C., Eriksson, J., Meadows, J.R.S., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. Nature *464*, 587–591.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., et al. (2013b). Genomic diversity and evolution of the head crest in the rock pigeon. Science *339*, 1063–1067.

Spencer, C.C. a, Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. *5*.

The International HapMap Consortium (2005). haplotype map of the human genome. Nature *437*, 1299–1320.

Tian, C., Gregersen, P.K., and Seldin, M.F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. Hum. Mol. Genet. *17*, R143–R150.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.

Visscher, P.M., Brown, M. a, McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Winkler, C. a., Nelson, G.W., and Smith, M.W. (2010). Admixture mapping comes of age *. Annu. Rev. Genomics Hum. Genet. *11*, 65–89.

Wray, G. a (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.

Zeggini, E., Rayner, W., Morris, A.P., Hattersley, A.T., Walker, M., Hitman, G.A., Deloukas, P., Cardon, L.R., and McCarthy, M.I. (2005). An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat. Genet. *37*, 1320–1322.

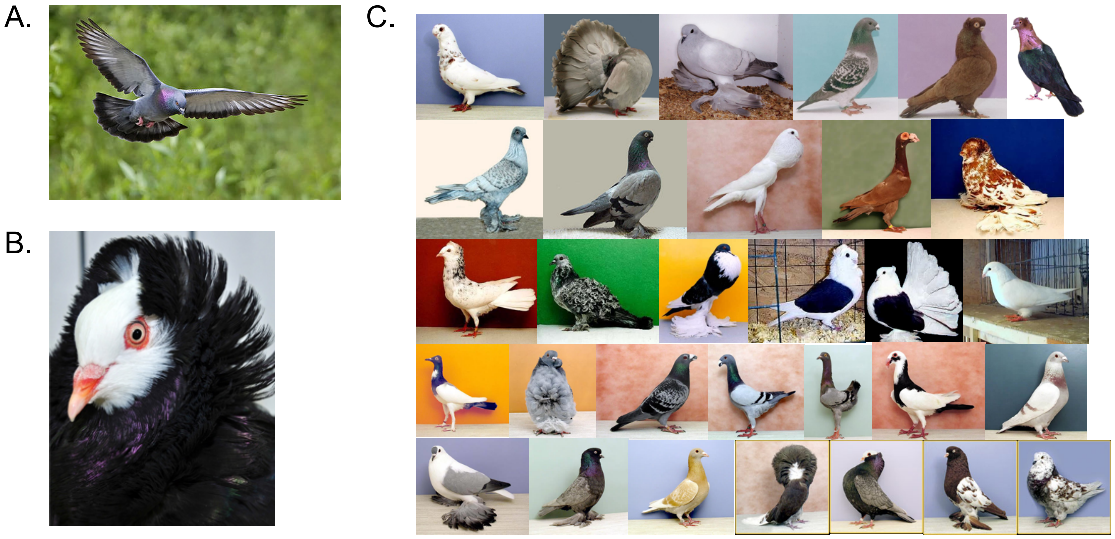
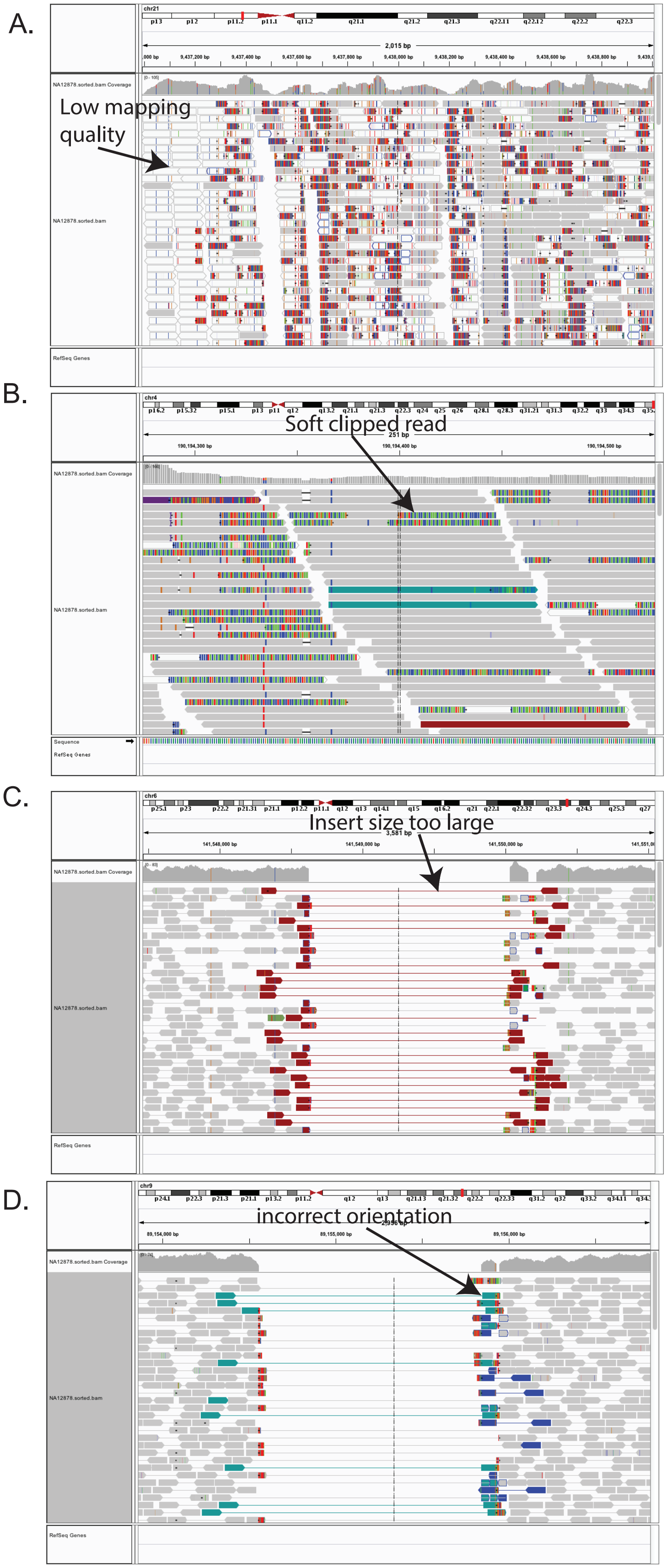


Figure 1.1. The phenotypic diversity amongst domesticated pigeons. A) The feral rock pigeon, *Columba livia*, is the progenitor of domestic pigeons. B) Mane crested Old Dutch Capuchin. The presence or absence of a headcrest is an example of a simple or binary trait. C) A compilation of different pigeon breeds exhibiting many derived phenotypes.

Figure 1.2. Mapping errors and structural variation present in a high coverage re-sequenced genome (NA12878). From top to bottom in each panel: the genomic position track, the smoothed read depth track, and the HTS paired-end read alignments. The intensity of the grey hue denotes the mapping quality of each read. Reads that are white have the lowest mapping quality. Reads that are not grey scaled have mates that map to other chromosomes. Reads with partial matches (soft clipped) are shown as grey bars with colored patches at the 5' or 3' end(s) of the read. A) Several microsatellites close to a centromere. Most reads within this region have low mapping qualities (white) and many are soft clipped (colored regions at the start or end of a read). B) Simple (CA)n repeat rich region near a telomere. The reads in the region have high mapping qualities and many are soft clipped. C) A homozygous partial deletion of a LINE-1 element. Reads are shown as mate-pairs. Read pairs shown in red are mapped too far apart, indicating a deletion, similarly reads at the breakpoint of the deletion are soft clipped. D) An inversion. Teal read pairs are reads that both map to the same strand, which is indicative of an inversion.


A.
chr21
2,015 bp
NA12878.sorted.bam Coverage
Low mapping quality
NA12878.sorted.bam
RefSeq Genes
B.
chr4
Soft clipped read
251 bp
190,194,300 bp
190,194,400 bp
190,194,500 bp
NA12878.sorted.bam Coverage
NA12878.sorted.bam
Sequence
RefSeq Genes
C.
chr6
Insert size too large
3,581 bp
141,548,000 bp
141,549,000 bp
141,550,000 bp
NA12878.sorted.bam Coverage
NA12878.sorted.bam
RefSeq Genes
D.
chr9
incorrect orientation
2,936 bp
89,154,000 bp
89,155,000 bp
89,156,000 bp
NA12878.sorted.bam Coverage
NA12878.sorted.bam
RefSeq Genes

CHAPTER 2

THE GENOTYPE PHENOTYPE ASSOCIATION TOOLKIT

2.1 Background

The field of population genetics has greatly benefited from the precipitous drop in the cost of high throughput sequencing (HTS). The bottleneck in the field has shifted from manually gathering genetic markers to processing population-level HTS data (Pool et al., 2010). While the theoretical foundation of population genetics is relatively stable, the HTS data used are rapidly evolving and growing in size. Traditional population genetic metrics were not designed to cope with missing genotypes or genotyping errors, frequent in all HTS data (Pompanon et al., 2005). If downstream analyses do not account for these errors, basic population genetic parameters can be biased. For example, HTS data can skew estimates of both the allele frequency spectrum and genetic diversity (Nielsen et al., 2011; Sackton et al., 2009). These estimates are skewed as HTS errors result in an abundance of rare variants. HTS errors can also affect the outcome of genotype-phenotype association (GPA) studies. The statistical models used in association studies rely on both accurate genotypes and estimates of the allele frequency spectrum. The methodology for population genetic and GPA analyses need to be updated to cope with the typical errors of HTS data. In response to the growing number of HTS population genetics and GPA studies, dozens of toolkits have been developed, each filling a specific niche.

One tool, ANGSD, focuses on estimating the allele frequency utilizing genotype likelihoods (Korneliussen, 2014). Another tool, Lositan, detects F_{st} outliers, a measure of population differentiation (Antao et al., 2008). Over the past several years, the software ecosystem has become crowded. Some of the options researchers have to choose from include VCFtools, ngsTools, PoPoolation, Pegas, Arlequin 3.0, and PlinkSEQ (Danecek et al., 2011; Excoffier et al., 2005; Fumagalli et al., 2014; Kofler et al., 2011; Paradis, 2010). While there are many published population genetic / GPA toolkits, not all are created equal. Some tools are difficult to install because of extensive external dependencies, while others are difficult to use because they have their own unique file formats. Also, many of these tools do not account for genotype uncertainty.

Recently, the benefits of accounting for genotype uncertainty when analyzing HTS data have become apparent (Fumagalli et al., 2013, 2014; Korneliussen, 2014; Li, 2011a). The uncertainty of a genotype call is quantified with genotype likelihoods. Formally, genotype likelihoods summarize the probability of the data (HTS reads at a position in the genome) given a genotype (SNPs, indels) (Li, 2011a; McKenna et al., 2010; Nielsen et al., 2011, 2012). In the case of a bi-allelic SNP, three genotype likelihoods will be reported by a variant caller: homozygous reference (0/1; 0), heterozygous (0/1; 1), and homozygous nonreference (1/1; 2). Genotype likelihoods are affected by mapping quality, base quality, and priors such as the allele frequency of the variant in the population. Tools that make use of genotype likelihoods have been shown, through simulation, to lower the mean bias between estimated and simulated population genetic parameters (Fumagalli et al., 2013). As many real-word datasets contain large numbers of individuals sequenced to low depth, it is critical that HTS analyses account for the uncertainty in the data by using genotype likelihoods (Abecasis et al., 2012; Kim et al., 2010; Li et al., 2011).

Here we present the Genotype Phenotype Association Toolkit, or GPAT++ for short. GPAT++ is a C++ population genetics toolkit that focuses on association and population differentiation testing. The algorithms in GPAT++ either directly use genotype likelihoods to account for HTS errors or provide sensible filters to facilitate rapid and reliable analyses. The tools within GPAT++ were developed to fit the needs of several unique genotype-phenotype association (GPA) study designs and later generalized into a user-friendly toolkit. Novel and traditional methods within GPAT++ have been vetted in eukaryotic systems (Domyan et al., 2014; Galati et al., 2014) and viral data. The main GPA test within GPAT++ is pFst, a single-marker test that has proven effective in mapping the genetic basis of traits in genotypic and pooled (allele frequency-based) HTS data. Besides the association tests and population differentiation statistics, GPAT++ also provides scripts for postprocessing of results and data visualization. Here we describe the novel methods implemented in GPAT++ and provide several examples of how to use GPAT++ on real data.

2.2 Methods

GPAT++ analyses start directly downstream of variant calling. GPAT++ takes a joint-called VCF formatted file containing population-level bi-allelic SNPs and INDELs (Danecek et al., 2011). This VCF file is compressed and indexed by chromosome and position using Tabix, which allows for range queries within GPAT++ (Li, 2011b). GPAT++ analyses prevent file splitting and merging by allowing the user to subset individuals within the VCF file. Care has been taken to homogenize GPAT++ command line options, reducing the amount of time required to learn GPAT++ workflows. A full list of analyses available in GPAT++ is presented in Table 2.1. GPAT++ provides a mix of modified and traditional GPA and population genetic measures. Most of our efforts have been directed towards our keystone tool, pFst.

2.2.1 Genotype phenotype association with a likelihood ratio test

GPAT++'s pFst test quantifies the difference between the target and background allele frequencies using a likelihood ratio test (LRT) under a binomial likelihood model. The basic LRT used within pFst has been widely adopted for association studies (Kim et al., 2010; Li, 2011a; Yandell et al., 2011). The null model of pFst LRT assumes that the allele frequencies of both the target (AF_T) and background (AF_B) groups are the same (same distribution), while the alternative hypothesis is that the allele frequencies of the two groups come from two separate distributions. The allelic counts in the model come from the genotype calls.

$$D \ = -2 \ * \ ln(\frac{B(N_C,K_C,AF_C)}{B(\ N_T,K_T,AF_T\ ) \times B(N_B,K_B,AF_B)}) \tag{2.1}$$

The binomial density function (B*(n, k, p*)) is parameterized by the number of successes *n*, the number of trials *k*, and the probability of success *p*. In the current application, *n* is the number of nonreference alleles in the target (N_T), background (N_B), and the target/background combined (N_C). The parameter *k* is the number of alleles in the target (K_T), background (K_B), and the target/background combined (K_C). The probability of success *p* corresponds to the target (AF_T), background (AF_B), and combined (AF_C) allele frequencies. The parameter *D* is the likelihood ratio test statistic. Larger *D* values can indicate that the null hypothesis should be rejected under the assumptions of the binomial model. A chi-sq lookup can be used to convert the *D* statistic into a p-value.

By default, pFst uses a modified form of the likelihood ratio, presented in equation 2.1, that incorporates the genotype likelihood information. The allele counts (equation 2.2 and 2.3) and the allele frequencies (equation 2.4) are estimated by integrating over the genotype likelihoods:

$$alleleA_{count\ estimate} = \sum_{i=0}^{i=n} (2 * gl_1 + gl_2) \tag{2.2}$$

$$alleleB_{count\ estimate} = \sum_{i=0}^{i=n} (2 * gl_3 + gl_2) \tag{2.3}$$

$$AF_{non-ref} = \frac{alleleB_{count\ estimate}}{alleleB_{count\ estimate} + alleleA_{count\ estimate}} \tag{2.4}$$

The variable *gl* represents the genotype likelihood and the subscript is the genotype (1,2,3; 0/0, 0/1, 1/1). The summation is over the subpopulations (cases and controls). The estimates from equations 2.2, 2.3, and 2.4 are used in equation 2.1.

For pooled samples the binomial equation used in 2.1 is replaced with a beta distribution. The parameters of the beta distributions are estimated using the methods of moments. The pooled version of pFst requires that both the target and background have more than one biological sample.

2.3 Applications of GPAT++

In this section, we demonstrate GPAT++'s functionality on HTS pigeon variant calls by examining the genetic basis of "headcrest", shown in Figure 2.1A (Shapiro et al., 2013). The presence or absence of headcrest can be attributed to a single allele (*cr*) of the *Ephb2* gene. *Cr* is a G to T nucleotide substitution causing a missense mutation (arginine to cysteine) in the EphB2 protein (Shapiro et al., 2013). In the head crest example, we use GPAT++ to identify the *cr* allele and interrogate the haplotype structure around *cr*. The GPAT++ commands and data used for the analyses are annotated in Table 2.2. GPAT++'s plotting scripts directly generated the figures presented in this section.

We begin be examining the association signal on scaffold612 between birds with and without head crest. Figure 2.1B-D shows the results of running of pFst, wcFst (Weir and Cockerham's F_{st}), and a Bayesian implementation of F_{st}. The highest peak for pFst and wcFst is *cr* at position 596,613. Weir and Cockerham's F_{st} and pFst generate association signals that are similar in appearance for scaffold612, while the Bayesian method appears to have more noise. One advantage of the Bayesian method over the other two methods is it provides confidence intervals around the best estimate of F_{st}. However, the Bayesian method is computationally burdensome, as each calculation requires thousands of Markov Chain Monte Carlo iterations. All three methods show elevated scores around the *cr* allele, suggesting that *cr* is an extended haplotype.

To explore the *cr* haplotype, we used GPAT++'s haplotype plotting and linkage disequilibrium (LD) tools. The core haplotype carrying the *cr* allele is ~10Kb long. We used GPAT++ to plot the haplotype within 25kb window of the cr allele (Figure 2.2A). At this distance, the haplotype sharing has completely decayed, although many of the haplotypes are very similar. Towards both edges of the haplotype plot, the haplotype diversity starts to increase due to recombination events between the *cr* haplotype and non-*cr* haplotypes. To quantify the amount of recombination in the region, we next used GPAT++'s linkage disequilibrium measure. GPAT measures LD as the D statistic (Devlin and Risch, 1995). Using this traditional method, we found little linkage disequilibrium around the *cr* allele for the birds with head crests (Figure 2.2B-C).

This is not unexpected as *cr* is recessive (fixed within the target population) and "D" cannot assay fixed alleles. The low density of assayable sites, visible in Figure 2.2B, is due to the fixed *cr* allele. To overcome this inherent shortcoming of "D", we allow users to provided external estimates of allele frequencies, allowing for fixed sites within a population to be assayed. By providing GPAT++'s LD method a background set of individuals we recovered the linkage disequilibrium surrounding the *cr* allele (Figure 2.2C).

2.4 Conclusion

GPAT++ provides powerful methodology for analyzing, summarizing, and visualizing complex population level datasets. Through the applied example, we demonstrate that GPAT++ is applicable to real-world datasets. The examples provided here are not all encompassing of GPAT++'s functionality, but merely a brief tutorial. For full documentation and additional examples of using GPAT++, please see the wiki: https://github.com/jewmanchue/vcflib/wiki. Continual development will focus on supporting new file formats and implementing additional population genetics metrics.

2.5 References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M. a, Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G. a (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A., and Luikart, G. (2008). LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics *9*, 323.

Danecek, P., Auton, A., Abecasis, G., Albers, C. a, Banks, E., DePristo, M. a, Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Devlin, B., and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics *29*, 311–322.

Domyan, E.T., Guernsey, M.W., Kronenberg, Z., Krishnan, S., Boissy, R.E., Vickrey, A.I., Rodgers, C., Cassidy, P., Leachman, S. a, Fondon, J.W., et al. (2014). Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr. Biol. *24*, 459–464.

Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online *1*, 47–50.

Fumagalli, M., Vieira, F.G., Korneliussen, T.S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. Genetics *195*, 979–992.

Fumagalli, M., Vieira, F.G., Linderoth, T., and Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. Bioinformatics *30*, 1486–1487.

Galati, D.F., Bonney, S., Kronenberg, Z., Clarissa, C., Yandell, M., Elde, N.C., Jerka-Dziadosz, M., Giddings, T.H., Frankel, J., and Pearson, C.G. (2014). DisAp-dependent striated fiber elongation is required to organize ciliary arrays. J. Cell Biol. *207*, 705–715.

Holsinger, K.E., Lewis, P.O., and Dey, D. (2002). A Bayesian approach to inferring population structure from docimant markers. Mol. Ecol. *11*, 1157–1164.

Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. Genet. Epidemiol. *34*, 479–491.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., and Schlötterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One *6*, e15925.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics *15*, 356.

Li, H. (2011a). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2011b). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. *21*, 940–951.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. *76*, 5269–5273.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One *7*, e37558.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics *26*, 419–420.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. *6*, 847–859.

Pool, J., Hellmann, I., Jensen, J., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. Genome Res. *20*, 291–300.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

Sackton, T.B., Kulathinal, R.J., Bergman, C.M., Quinlan, A.R., Dopman, E.B., Carneiro, M., Marth, G.T., Hartl, D.L., and Clark, A.G. (2009). Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. Genome Biol. Evol. *1*, 449–465.

Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., et al. (2013b). Genomic diversity and evolution of the head crest in the rock pigeon. Science (80-.). *339*, 1063–1067.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution (N. Y). 1358–1370.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.

Table 2.1 Primary analyses available in GPAT++. The last two columns state the requirements of each method: genotype data or pooled sequencing or phased or unphased diploids.

Method	Genotype or Pooled	Reference	Phased Variant requirement
Population statistics	Both	NA	Phased/unphased
pFst	Both	NA	Phased/unphased
Weir and Cockerham Fst	Genotype only	(Weir and Cockerham, 1984)	Phased/unphased
Bayesian Fst	Genotype only	(Holsinger et al., 2002)	Phased/unphased
ABBA-BABA	Genotype only	(Durand et al., 2011)	Phased/unphased
iHS	Genotype only	(Voight et al., 2006)	Phased only
XP-EHH	Genotype only	(Sabeti et al., 2007)	Phased only
Linkage disequilibrium (D)	Genotype only	(Devlin and Risch, 1995)	Phased only
Nucleotide diversity	Genotype only	(Nei and Li, 1979)	Phased only

Table 2.2 GPAT++ commands used for the head crest analyses

Task: Identify genetic variants with different allele frequencies (associated with head crest). **Command:** ../bin/pFst --target 1,20,25,29,30,38,43,46 --background 2,3,4,5,6,7,21,22,22,23,24,26,26,28,31,32,33,34,35,36,37,39,40,41,42,44,45 --deltaaf 0.0 --file scaffold612.vcf --counts --type PL > 612.counts 2> 612.err **Explanation:** A list of target and background birds is provided to pFst using the –target and –background flags. Genetic variants with allele frequency differences less than –deltaaf are skipped. The –counts flag instructs pFst to use the genotype count model. The –type flag denotes the format of the genotype likelihoods in the VCF. The STDOUT and STDERR are redirected to 612.counts and 612.err, respectively.
Task: Plot the output of pFst **Command:** R --vanilla < ../bin/plotPfst.R --args 612.counts **Explanation:** GPAT++ analyses use the statistical language R for plotting functions. The R script (plotPfst.R) is passed the pFst file as an argument. The R script writes a plot file with the same prefix as the input file and the suffix is the time and date.
Task: Identify genetic variants that show evidence of differential inbreeding between populations using Weir and Cockerham's F_{st}. For this example the two populations are birds with and without head crests. **Command:** bin/wcFst --target 1,20,25,29,30,38,43,46 --background 2,3,4,5,6,7,21,22,22,23,24,26,26,28,31,32,33,34,35,36,37,39,40,41,42,44,45 --deltaaf 0.0 –type PL --file scaffold612.vcf > 612.wcfst.txt 2> 612.wcfst.err **Explanation:** The command line options are identical to the pFst command.
Task: Plotting the output of wcFst. **Command:** R --vanilla < ../bin/plotWCfst.R --args 612.wcfst.txt **Explanation:** The same syntax as plotPfst.

Table 2.2 continued

Task: Identify genetic variants that show evidence of differential inbreeding between populations using a Bayesian method. For this example the two populations are birds with and without head crests.

Command: bin/bFst --target 1,20,25,29,30,38,43,46 --background 2,3,4,5,6,7,21,22,22,23,24,26,26,28,31,32,33,34,35,36,37,39,40,41,42,44,45 --deltaaf 0.0 --file scaffold612.vcf > 612.bFst

Explanation: The command line options are identical to the pFst command.

Command: R --vanilla < ../bin/plotBfst.R --args 612.bFst

Explanation: The same syntax as plotPfst.

Task: Visualize the haplotypes around the *Ephb2* gene. This is a two-step process. The first step generates the input for the plotting script.

Command: bin/plotHaps --target 1,20,25,29,30,38,43,46 --file ~/gpat/samples/scaffold612.phased.vcf.gz --type GP --region scaffold612:584000-609000 > headCrest.haps.txt

Explanation: Generating the input for the haplotype plotting script. –target are the individuals to plot. –type is the genotype format. In this case the VCF file is from Beagle 4.0. –region specifies the bounds of the analysis.

Task: Plot the haplotype data generated in the prior step.

Command: R --vanilla < bin/plotHaplotypes.R --args headCrest.haps.txt

Explanation: Plot the haplotypes using the output of the last command.

Task: Calculated the linkage disequilibrium between genetic variants for the birds with a head crest.

Command: bin/LD --target 1,20,25,29,30,38,43,46 –d –w 20 > within-pop.ld.11.txt

Explanation: Calculate the linkage average linkage disequilibrium within 20 SNP windows. –target is the individuals to use. –d is the flag instructs the program to use the non-reference haplotype frequency. –w is the SNP window size.

Table 2.2 continued

Task: Plotting the LD calculated in the previous step. **Command:** R --vanilla < bin/plotLD.R --args within-pop.ld.11.txt **Explanation:** Plotting mean LD using the output from the last command
Task: Calculating the LD around the Ephb2 for birds with a head crest. Unlike the previous LD calculation a background will be used. **Command:** bin/LD --target 1,20,25,29,30,38,43,46 --background 2,3,4,5,6,7,21,22,22,23,24,26,26,28,31,32,33,34,35,36,37,39,40,41,42,44,45 --type GP --file samples/scaffold612.phased.vcf -e -d -w 20 > between-pop.ld.11.txt **Explanation:** Calculating the LD for the target population (--target) using the background individuals for the haplotype frequency expectation. –e is the flag that instructs LD to use external allele frequencies.
Task: Plotting the LD calculated in the previous step. **Command:** R --vanilla < bin/plotLD.R --args between-pop.ld.11.txt **Explanation:** Plotting mean LD using the output from the last command

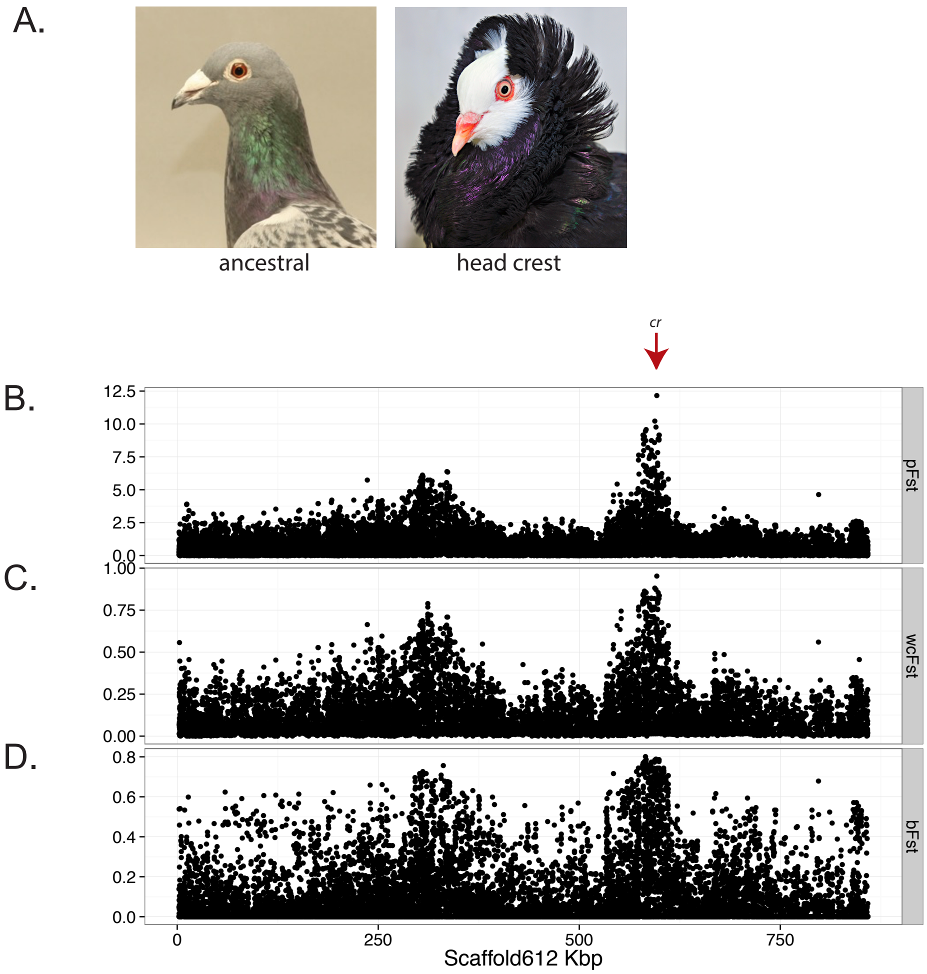
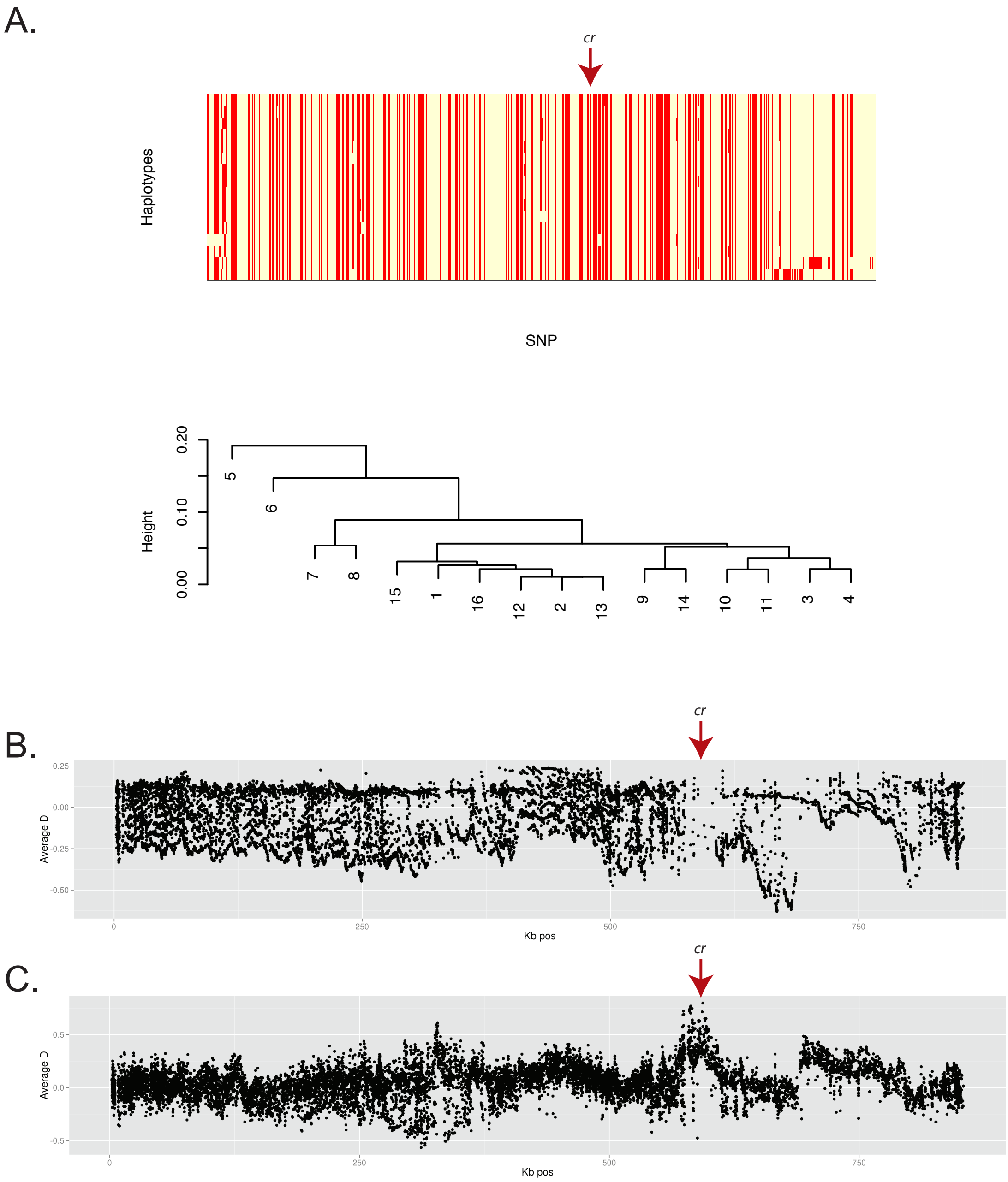


Figure 2.1. Mapping the genetic basis of headcrest using three different methods A) An uncrested and crested bird. B-D) The genotype-phenotype association scans for headcrest on scaffold612. The position of the headcrest allele (*cr*) is highlight by the arrow. B) pFst run with genotype counts. C) Weir and Cockerham's F_{st}. D) Bayesian F_{st}.

Figure 2.2. Analysis of the *cr* allele using GPAT++ haplotype-based methods. A) Haplotype plot of scaffold612 (584Kb-609Kb) for birds with the *cr* allele. Each row is a single haplotype and the columns are variable sites (bi-allelic SNPs and indels). The haplotypes are clustered by similarity. The dendrogram below the haplotype plot summarizes the clustering. The haplotypes carrying the *cr* allele are very similar, resulting in the “comb” like appearance in the dendrogram. B-C) The average Linkage disequilibrium (measured with the D statistic) across scaffold612 is plotted. The headcrest allele, *cr*, is notated with an arrow. B) The D static using the allele frequency and haplotype frequency within the target population. Many of the variable sites around *cr* cannot be assayed as they are fixed within the target population. C) By using the background population’s allele frequency as the expectation, we are able to score sites fixed in the target, but variable in the background.


A.
cr
Haplotypes
SNP
Height
0.20
0.10
0.00
5
6
7
8
15
1
16
12
2
13
9
14
10
11
3
4
B.
cr
Average D
0.25
0.00
-0.25
-0.50
0
250
500
750
Kb pos
C.
cr
Average D
0.5
0.0
-0.5
0
250
500
750
Kb pos

CHAPTER 3

GENOMIC DIVERSITY AND EVOLUTION OF THE HEAD CREST IN THE ROCK PIGEON

From Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., et al. (2013). Genomic diversity and evolution of the head crest in the rock pigeon. Science *339*, 1063–1067. reprinted with permission from AAAS

3.1 Contribution

For this paper, I carried Genotype Phenotype Association (GPA) for head crest (Figure 2 in the paper), estimating general population genetic parameters, inferring breed histories and the relationship between the 40 sequenced birds (Figure 1 in the paper). There were many technical challenges in this project. For example, the variant calls I received from our collaborators required filtering and care quality assurance. In total, it took a full year to do all of the analyses for this project.

This project encouraged me to write GPAT++, which expedites many of the analyses in the paper. The 40 sequenced birds have been an immense resource for developing new algorithms and testing methods already in existence. The regions around the head crest locus *Ephb2* has been used so extensively for GPAT++ development that it is distributed with the code for a tutorial.

enrichment, should be paced by changes in dust concentration. During TIII, the change in dust occurs earlier than the change in ice isotope at both EDC and Vostok (figs. S7 and S8), whereas these two records are approximately in phase during TI (fig. S8). This could explain why the Vostok δ^{40}Ar record is in advance with respect to the aCO_2 record, without contradicting our finding of synchronous changes in aCO_2 and AT. During TII at EDC (fig. S8), on the other hand, the change in δ^{15}N occurs at a deeper depth than the change in dust. Dust concentration therefore cannot be the only factor influencing the LID.

Our results are also in general agreement with a recent 0- to 400-year aCO_2-AT average lag estimate for TI (*20*), using a different approach. Although this study does not make any assumption about the convective zone thickness, it is based on coastal cores, which might be biased by local changes in ice sheet thickness; and firn densification models, which may not be valid for past conditions (see the supplementary materials for a more detailed discussion).

Our chronology and the resulting aCO_2-AT phasing strengthens the hypothesis that there was a close coupling between aCO_2 and AT on both orbital and millennial time scales. The aCO_2 rise could contribute to much of the AT change during TI, even at its onset, accounting for positive feedbacks and polar amplification (*21*), which magnify the impact of the relatively weak rCO_2 change (Fig. 4) that alone accounts for ~0.6°C of global warming during TI (*21*). Invoking changes in the strength of the Atlantic meridional overturning circulation is no longer required to explain the lead of AT over aCO_2 (*22*).

Given the importance of the Southern Ocean in carbon cycle processes (*23*), one should not exclude the possibility that aCO_2 and AT are interconnected through another common mechanism such as a relationship between sea ice cover and ocean stratification. Although the tight link between aCO_2 and AT suggests a major common mechanism, reviews of carbon cycle processes suggest a complex association of numerous independent mechanisms (*2*, *23*).

Changes in aCO_2 and AT were synchronous during TI within uncertainties. Our method, based on air ^{15}N measurements to determine the ice/gas depth shift, is currently being used in the construction of a common and optimized chronology for all Antarctic ice cores (*24*, *25*). The assumption that no convective zone existed at EDC during TI might be tested in the future by using Kr and Xe isotopes (*26*). Further studies on the firn are needed to understand the causes of the past variations of the LID, such as the possible impact of impurity concentrations on the densification velocity. Although our study was focused on the relative timing of TI climatic records extracted from Antarctic ice cores, there is now the need to build a global chronological framework for greenhouse gases, temperature reconstructions, and other climate proxies at various locations (*22*). Although the timings of the Bølling, Younger Dryas, and Holocene onsets as visible in the methane records are now well constrained by a layer-counted Greenland chronology (*27*), determining the timing of the onset of TI in Antarctic records remains challenging. Modeling studies using coupled carbon cycle–climate models will be needed to fully explore the implications of this synchronous change of AT and aCO_2 during TI in order to improve our understanding of natural climate change mechanisms.

References and Notes

1. E. Monnin *et al.*, *Science* **291**, 112 (2001).
2. A. Lourantou *et al.*, *Global Biogeochem. Cycles* **24**, GB2015 (2010).
3. H. Fischer, M. Wahlen, J. Smith, D. Mastroianni, B. Deck, *Science* **283**, 1712 (1999).
4. L. Loulergue *et al.*, *Clim. Past* **3**, 527 (2007).
5. F. Parrenin *et al.*, *Clim. Past* **8**, 1239 (2012).
6. C. Goujon, J.-M. Barnola, C. Ritz, *J. Geophys. Res.* **108**, ACL10/1-10 (2003).
7. F. Parrenin *et al.*, *Clim. Past* **3**, 243 (2007).
8. H. Craig, Y. Horibe, T. Sowers, *Science* **242**, 1675 (1988).
9. T. A. Sowers, M. Bender, D. Raynaud, Y. L. Korotkevich, *J. Geophys. Res.* **97**, 15683 (1992).
10. A. Landais *et al.*, *Quat. Sci. Rev.* **25**, 49 (2006).
11. G. B. Dreyfus *et al.*, *Quat. Sci. Rev.* **29**, 28 (2010).
12. J. P. Severinghaus *et al.*, *Earth Planet. Sci. Lett.* **293**, 359 (2010).
13. T. F. Stocker, S. J. Johnsen, *Paleoceanography* **18**, 1 (2003).
14. G. M. Raisbeck, F. Yiou, J. Jouzel, T. F. Stocker, *Clim. Past* **3**, 541 (2007).
15. S. Barker *et al.*, *Science* **334**, 347 (2011).
16. M. Hörhold *et al.*, *Earth Planet. Sci. Lett.* **325–326**, 93 (2012).
17. P. Köhler, G. Knorr, D. Buiron, A. Lourantou, J. Chappellaz, *Clim. Past* **7**, 473 (2011).
18. G. Myhre, E. J. Highwood, K. P. Shine, F. Stordal, *Geophys. Res. Lett.* **25**, 2715 (1998).
19. N. Caillon *et al.*, *Science* **299**, 1728 (2003).
20. J. B. Pedro, S. O. Rasmussen, T. D. van Ommen, *Clim. Past* **8**, 1213 (2012).
21. P. Köhler *et al.*, *Quat. Sci. Rev.* **29**, 129 (2010).
22. J. D. Shakun *et al.*, *Nature* **484**, 49 (2012).
23. H. Fischer *et al.*, *Quat. Sci. Rev.* **29**, 193 (2010).
24. L. Bazin *et al.*, *Clim. Past Discuss.* **8**, 5963 (2012).
25. D. Veres *et al.*, *Clim. Past Discuss.* **8**, 6011 (2012).
26. J. P. Severinghaus, A. Grachev, B. Luz, N. Caillon, *Geochim. Cosmochim. Acta* **67**, 325 (2003).
27. A. Svensson *et al.*, *Clim. Past* **4**, 47 (2008).
28. J. Jouzel *et al.*, *Science* **317**, 793 (2007).
29. L. Loulergue *et al.*, *Nature* **453**, 383 (2008).

Acknowledgments: We thank O. Watanabe, B. Stenni, and EPICA community members for giving access to, respectively, the DF1, TALDICE, and EDML isotopic data; L. Loulergue, D. Buiron, and T. Blunier for giving access to, respectively, the EDC-EDML, TALDICE, and GRIP CH_4 data; G. Dreyfus for giving access to the δ^{15}N isotopic data; and G. Delaygue, J. Chappellaz, S. Barker, and A. Ganopolski for helpful discussions. This work greatly benefited from constructive comments by two anonymous reviewers. This work had support from the French Agence Nationale de la Recherche (project ANR-07-BLAN-0125 "Dome A" and ANR-09-SYSC-001 "ADAGE"). This work is a contribution to EPICA, a joint European Science Foundation/European Commission scientific program, funded by the European Union and by national contributions from Belgium, Denmark, France, Germany, Italy, the Netherlands, Norway, Sweden, Switzerland, and the United Kingdom. Main logistical support was provided by the Institut Paul Emile Victor and the Programma Nazionale Ricerche in Antartide at Dome C. We thank the technical teams in the field and at the different laboratories. This is EPICA publication no. 291.

Supplementary Materials
www.sciencemag.org/cgi/content/full/339/6123/1060/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S7
Database S1

20 June 2012; accepted 9 January 2013
10.1126/science.1226368

Genomic Diversity and Evolution of the Head Crest in the Rock Pigeon

Michael D. Shapiro,[1]* Zev Kronenberg,[2] Cai Li,[3,4] Eric T. Domyan,[1] Hailin Pan,[3] Michael Campbell,[2] Hao Tan,[3] Chad D. Huff,[2,5] Haofu Hu,[3] Anna I. Vickrey,[1] Sandra C. A. Nielsen,[4] Sydney A. Stringham,[1] Hao Hu,[5] Eske Willerslev,[4] M. Thomas P. Gilbert,[4,6] Mark Yandell,[2] Guojie Zhang,[3] Jun Wang[3,7,8]*

The geographic origins of breeds and the genetic basis of variation within the widely distributed and phenotypically diverse domestic rock pigeon (*Columba livia*) remain largely unknown. We generated a rock pigeon reference genome and additional genome sequences representing domestic and feral populations. We found evidence for the origins of major breed groups in the Middle East and contributions from a racing breed to North American feral populations. We identified the gene *EphB2* as a strong candidate for the derived head crest phenotype shared by numerous breeds, an important trait in mate selection in many avian species. We also found evidence that this trait evolved just once and spread throughout the species, and that the crest originates early in development by the localized molecular reversal of feather bud polarity.

Since the initial domestication of the rock pigeon in Neolithic times (*1*), breeders have selected striking differences in behavior, vocalizations, skeletal morphology, feather ornaments, colors, and color patterns to establish over 350 breeds (*2*). In many cases, the number and magnitude of differences among breeds are more characteristic of macroevolutionary changes than of changes within a single species (*2*, *3*). Indeed, Charles Darwin was so fascinated by domestic pigeons that he repeatedly called attention to this dramatic example of diversity within a species to communicate his ideas about natural selection (*3*, *4*).

The genetic architecture for many derived traits in pigeons is probably relatively simple (*5*, *6*), probably more so than that for interspecific trait variation among many wild species, because breeders often focus on qualitative rather than quantitative variation; this increases the chance of identifying genes responsible for differences among breeds. Additionally, several morphological traits show similar patterns of variation in different breeds, making it possible to test whether the same or different genes underlie similar phenotypes. Despite these advantages, the pigeon is underused as a model for the molecular genetic basis of avian variation because of the paucity of genetic and genomic resources for this bird.

We examined genomic diversity, genetic structure, and phylogenetic relationships among domestic breeds and feral populations (free-living birds descended from escaped domestics) of the rock pigeon. The pigeon reference genome was sequenced from a male Danish tumbler with the Illumina HiSeq 2000 platform, and we also resequenced 40 additional *Columba livia* genomes to 8- to 26-fold coverage (38 individuals from 36 domestic breeds and two feral pigeons) (*7*). Genome-wide nucleotide diversity in the rock pigeon ($\pi = 3.6 \times 10^{-3}$) and the mutation rate estimate in the pigeon lineage (1.42×10^{-9} substitutions per site per year ± 2.60×10^{-12} SE) are comparable to those of other avian species (*8*, *9*). The observed heterozygosity indicates a large effective population size for the rock pigeon of $N_e \approx 521{,}000$; demographic inferences based on the allele frequency spectrum indicate that, aside from a very recent bottleneck, N_e has been remarkably stable over the past 1.5 million generations (*7*).

Patterns of linkage disequilibrium (LD) are indicative of haplotype sizes and genome-wide recombination rates and inform decisions about genetic mapping strategies. Using genotype data from the 40 resequenced *C. livia* genomes, we

[1]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA. [2]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA. [3]BGI–Shenzhen, Shenzhen, 518083, China. [4]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark. [5]Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA. [6]Ancient DNA Laboratory, Murdoch University, Perth, Western Australia 6150, Australia. [7]Department of Biology, University of Copenhagen, DK-1165 Copenhagen, Denmark. [8]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-1165 Copenhagen, Denmark.

*To whom correspondence should be addressed. E-mail: mike.shapiro@utah.edu (M.D.S.); wangj@genomics.org.cn (J.W.)

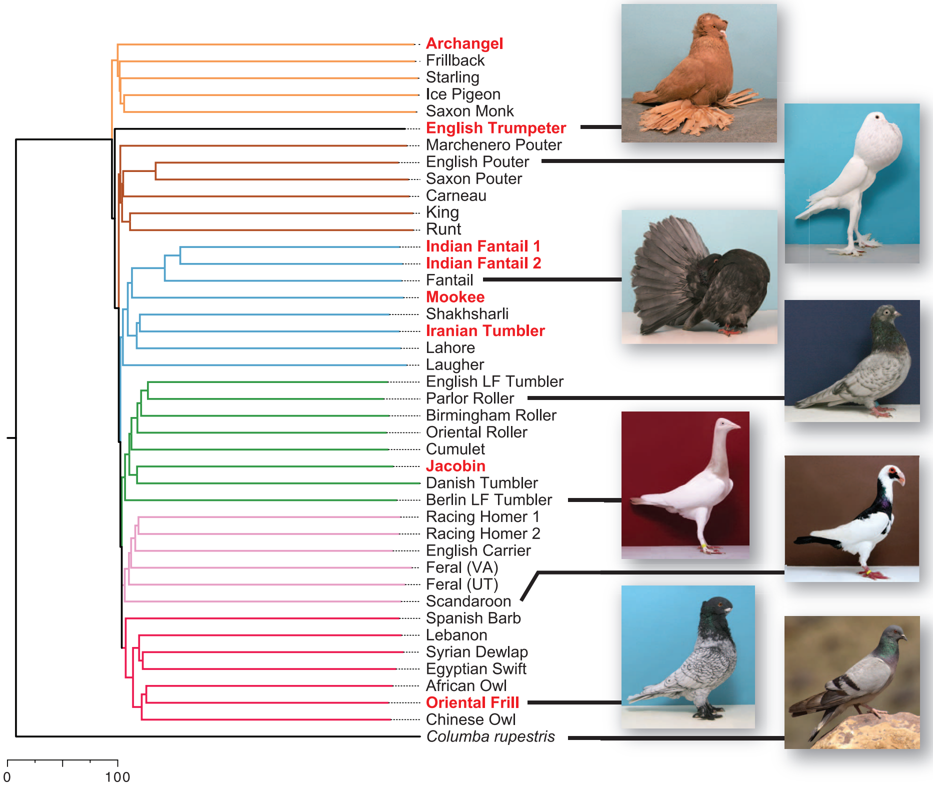


Fig. 1. Relationships among rock pigeons and the hill pigeon *C. rupestris*. A consensus neighbor-joining tree based on 1.48 million genomic SNPs and 1000 bootstrap replicates (see fig. S16 for bootstrap support) is shown. Branches are colored according to traditional breed groups (*12*) and/or geographic affinities: orange, toy breeds; brown, pouters and utility breeds; light blue, Indian and Iranian breeds; green, tumblers and highflyers; pink, homers and wattle breeds; red, Mediterranean and owl breeds; black, voice characteristics (*14*). Bold red lettering indicates breeds with the head crest phenotype. Scale bar, Euclidean distance. [Photo credits: T. Hellmann (domestic breeds) and M. V. Shreeram (*C. rupestris*)]

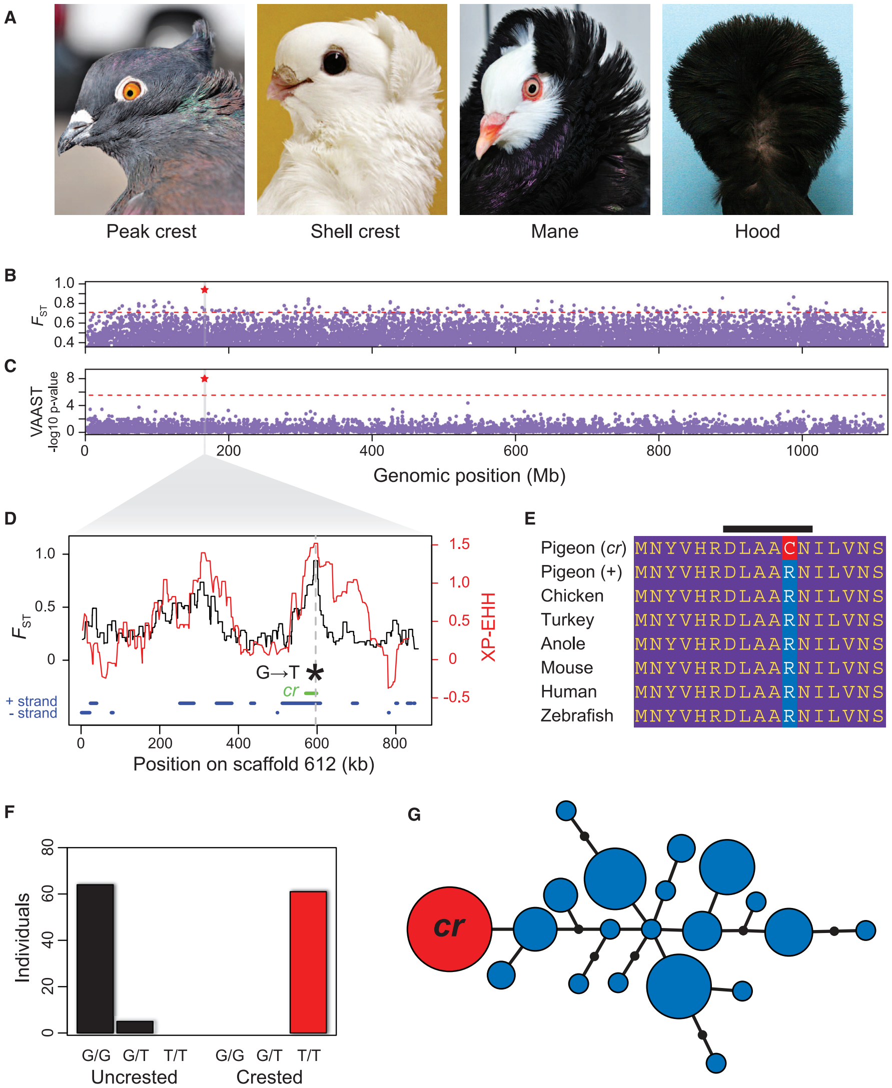


Fig. 2. *EphB2* is associated with the derived head crest phenotype. (**A**) Head crests are variable among breeds (left to right: Indian fantail, Old German owl, Old Dutch capuchin, Jacobin). (**B**) F_{ST} between crested and uncrested pigeons, with maximum value for individual SNPs plotted for nonoverlapping 100-kb windows across the genome. Red star, window with the highest score. Dashed red line, top 1% of scores. (**C**) Genome-wide VAAST scan. Each dot represents a single gene. Red star, gene with the highest score. Dashed red line, genome-wide significance cutoff. (**D**) Magnification of scaffold 612 in shaded region of (B) and (C). Black trace, maximum F_{ST} between crested and uncrested birds over a 300-SNP window. Red trace, unstandardized cross-population extended haplotype homozygosity (XP-EHH); higher values are evidence of selection (see fig. S21, genome-wide plot). Dashed vertical line, position of the lone genome-wide significant VAAST hit. Green bar, the 27.4-kb haplotype shared by all crested birds, includes only the *EphB2* gene. Blue bars, gene predictions on + and – DNA strands. (**E**) The *cr* mutation induces a charge-changing amino acid substitution; black bar, highly conserved DLAARN motif of catalytic loop. (**F**) Genotypes of 159 birds from 79 breeds at the *cr* locus are perfectly associated with the crest phenotype under a recessive model. (**G**) Network diagram of the minimal 11-kb haplotype shared by all resequenced rock pigeons with the *cr* mutation (also see fig. S23). Many haplotypes contain the + allele (blue), but only one contains the *cr* SNP (red). The sizes of the circles are proportional to the number of chromosomes containing a haplotype. Line segments represent single-nucleotide differences. [Jacobin photo credit: T. Hellmann]

found that mean "useful LD" (10) (coefficient of determination, $r^2 > 0.3$) decays in 2.2 kb (fig. S10J). This suggests that we should expect little LD between typical pairs of genes in an analysis across breeds; thus, the pigeon is well suited for association-mapping strategies.

We leveraged our whole-genome data to determine breed relationships, using 1.48 million variable loci. A neighbor-joining tree rooted on *C. rupestris*, the sister species of *C. livia* (11), yielded several well-supported groups (Fig. 1 and fig. S16). Notably, the two feral pigeons grouped with the wattle and homer breeds (Fig. 1, pink branches), supporting the idea that escaped racing homers are probably major contributors to feral populations (12). As with many domesticated species, pigeon evolution is probably not exclusively linear or hierarchical (12). We therefore examined genetic structure among breeds by analyzing 3950 loci with ADMIXTURE (13) and found a best model fit at $K = 1$ (a single population, where K is the number of assumed ancestral populations). However, higher values of K can also be biologically informative (figs. S17 to S20). Our analysis includes some of the oldest lineages of domestic pigeons and breeds that were not exported from the Middle East until the late 19th or early 20th centuries (14), providing information about likely geographic origins of breeds and their exchange along ancient trade routes (7).

Derived traits in domesticated birds tend to evolve along a predictable temporal trajectory, with color variation appearing in the earliest stages of domestication, followed by plumage and structural (skeletal and soft tissue) variation, and finally behavioral differences (2). One of the genetically simplest derived traits of pigeons is the head crest. Head crests are common ornaments in many bird species (2) and are important display structures in mate selection (15). In pigeons, head crests consist of neck and occipital feathers with reversed growth polarity, so that the feathers grow toward the top of the head instead of down the neck. Crests can be as small and simple as a peak of feathers or as elaborate as the hood of the Jacobin, which envelops the head (Fig. 2A). Classical genetics experiments suggest that the head crest segregates as a simple Mendelian recessive trait (6, 14). Moreover, previous studies suggest that the same locus controls the presence of a crest in numerous breeds, either with alternative alleles at this locus or additional modifier loci controlling the extent of crest development (6, 14).

We resequenced eight individuals with head crests to directly test whether the same mutation controls crest development in different breeds. We sorted genomic variants from birds with and without head crests into separate bins and calculated allele frequency differentiation (F_{ST}) across the genome (Fig. 2B). We identified a region of high differentiation between crested and uncrested birds in the pigeon ortholog of *Ephrin receptor B2* (*EphB2*; $F_{ST} = 0.94$, top hit genome-wide; fig. S22A) (Fig. 2D). The role of *EphB2* in feather growth is not known, but it plays important roles in tissue patterning and morphogenesis and is a member of a receptor tyrosine kinase family that mediates development of the feather cytoskeleton (16, 17). All eight crested birds were homozygous for a T nucleotide at scaffold 612, position 596613 (hereafter, the *cr* allele), whereas uncrested birds were heterozygous ($n = 3$) or homozygous ($n = 30$, including the uncrested outgroup *C. rupestris*) for the putatively ancestral C nucleotide (the + allele). These results were consistent with the known simple recessive architecture of the trait and implicated a common polymorphism associated with head crest development in multiple breeds with different genetic histories (Fig. 1). This trend extended well beyond our resequencing panel: We genotyped an additional 61 crested birds from 22 breeds and 69 uncrested birds from 57 breeds, and found a perfect association between *cr/cr* genotype and the crest phenotype (Fig. 2F). By treating the genomes of crested and uncrested birds as separate populations, we also found suggestive evidence for positive selection around the *cr* allele using cross-population extended haplotype homozygosity analysis (Fig. 2D and figs. S21 and S22B).

We then used the Variant Annotation, Analysis, and Search Tool [VAAST (18)] to investigate the pigeon genomes for additional coding changes associated with the head crest phenotype. This identified one gene with genome-wide significance: *EphB2*, and specifically the *cr* single-nucleotide polymorphism (SNP) ($P_{genome} = 2.0 \times 10^{-8}$) (Fig. 2, C and D). The *cr* allele has a predicted charge-changing arginine (basic) to cysteine (polar uncharged) transition in the catalytic loop of the intracellular tyrosine kinase domain of *EphB2* (Fig. 2E). This amino acid position is invariant among other vertebrates, suggesting strong purifying selection for conserved protein function. The same DLAARN to DLAACN motif change we observe in EphB2 is sufficient to abrogate kinase activity in human and mouse orthologs of the protein tyrosine kinase ZAP-70, and in both mammals and pigeons the mutant phenotypes are inherited recessively (19). Hence, the pigeon *cr* mutation probably abrogates kinase activity in EphB2 and disrupts downstream signal propagation, consistent with the high VAAST score for this gene. *EphB2* is therefore a convincing candidate for the *cr* locus of classical pigeon genetics (5–7, 14).

In several wild and domesticated species, the repeated evolution of a derived trait has occurred by selection on the same gene, possibly owing to the repeated selection on the same allele or haplotype (20–22). Similarly, the *cr* SNP is part of a 27.4-kb haplotype that is shared by all crested pigeons, suggesting that the mutation occurred just once and spread to multiple breeds by introgression among domestic breeds, or was selected repeatedly from a standing variant in wild rock pigeons (Fig. 2G and fig. S23; the core haplotype containing the *cr* mutation is reduced to 11 kb when uncrested heterozygotes are included). The only gene present in the shared *cr* haplotype is *EphB2* (Fig. 2D, green bar), although at this time we cannot rule out the presence of regulatory variants that might alter the expression of another gene. Crested members of the toy, fantail, Iranian, Jacobin, and owl breed groups are not more closely related to each other than to uncrested breeds (Fig. 1). Nevertheless, members of these groups had head crests

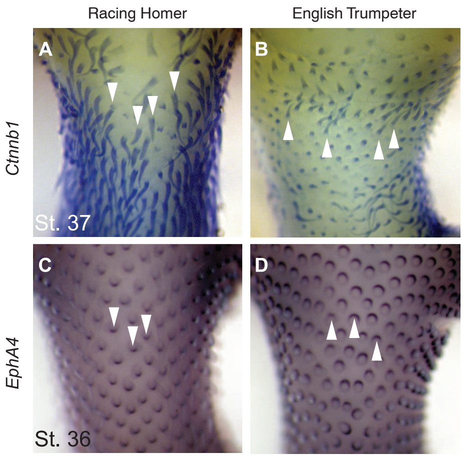


Fig. 3. Feather bud polarity is reversed in the *cr* mutant. (**A** and **B**) Expression of the feather structural gene *Ctnnb1* reveals the direction of outgrowth of early feather buds. St., Hamburger-Hamilton embryonic stage. (A) Neck and occipital head expression of *Ctnnb1* in an embryo of the uncrested racing homer. Feather buds point downward along the contour of the head and neck (arrowheads). (B) Occipital feather buds point upward in the equivalent region of the crested English trumpeter, indicating morphological reversal of feather orientation. (**C** and **D**) Expression of the polarity marker *EphA4* was assayed at an earlier developmental stage to test whether feather placodes, the ectodermal thickenings that give rise to feather buds, are also reversed. (C) Polarity marker *EphA4* is expressed posteriorly (arrowheads) in feather placodes of the racing homer. (D) The polarity of placodes is reversed in the English trumpeter. Expression of *EphB2* in the skin is weak and unpolarized at this stage in both morphs (fig. S26).

hundreds of years ago (*14*), so some of these introgression events must have occurred in the distant past. Breeds with a wide variety of crest phenotypes share the same derived allele; therefore, allelic variation at the *cr* locus alone does not control all aspects of crest development (*14*). Other genetic and developmental factors beyond this locus must contribute to variation in crest morphology, akin to the presumed complex genetic architecture of species-level divergence in feather ornaments (*2*).

In crested pigeons, feather placode polarity and bud outgrowth are inverted during embryogenesis (Fig. 3). Expression of *EphB2* is not polarized in early placodes (fig. S26), so the effects of the *cr* mutation on feather polarity are probably exerted earlier in development. Why might the crest phenotype be limited to the head and neck? In Naked neck chicken mutants, regionalized production of retinoic acid allows uniform up-regulation of *Bmp7* expression to change skin phenotypes in the neck but not the body (*23*). Similarly, the head crests of several chicken breeds, in which feathers are elongated but do not have a reversed growth trajectory as in pigeons, are localized to the top of the head, probably due to ectopic expression of *Hox* positional cues (*24*). Together these examples provide evidence for regionalization of the developing head and neck skin in the chicken. We propose that analogous mechanisms might underlie skin regionalization in the pigeon and allow *cr* to change feather polarity in the occiput and neck, but not elsewhere.

Our study of domestic rock pigeons illustrates how combining comparative genomics and population-based analyses forwards our understanding of genetic relationships and the genomic basis of traits. Many of the traits that vary among pigeon breeds also vary among wild species of birds and other animals (*2*, *25*); thus, pigeons are a model for identifying the genetic basis of variation in traits of general interest. Moreover, variation in many traits in domestic pigeons, including the head crest phenotype described here, is constructive rather than regressive: Breeds derived from the ancestral rock pigeon possess traits that the ancestor does not have. Although adaptive regressive traits are important, the genetic basis of constructive traits in vertebrates remains comparatively poorly understood. The domestic pigeon is thus a promising model with which to explore the genetic architecture of derived, constructive phenotypes in a bird that is amenable to genetic, genomic, and developmental investigation.

References and Notes

1. C. A. Driscoll, D. W. Macdonald, S. J. O'Brien, *Proc. Natl. Acad. Sci. U.S.A.* **106** (suppl. 1), 9971 (2009).
2. T. D. Price, *Genetica* **116**, 311 (2002).
3. C. Darwin, *On the Origin of Species by Means of Natural Selection* (John Murray, London, 1859).
4. C. Darwin, *The Variation of Animals and Plants Under Domestication* (John Murray, London, 1868), vol. 1.
5. T. H. Morgan, *Biol. Bull.* **21**, 215 (1911).
6. A. Sell, *Breeding and Inheritance in Pigeons* (Schober Verlags-GmbH, Hengersberg, Germany, 1994).
7. See the supplementary materials on *Science* Online.
8. C. N. Balakrishnan, S. V. Edwards, *Genetics* **181**, 645 (2009).
9. H. Ellegren *et al.*, *Nature* **491**, 756 (2012).
10. J. Aerts *et al.*, *Cytogenet. Genome Res.* **117**, 338 (2007).
11. K. P. Johnson *et al.*, *Auk* **118**, 874 (2001).
12. S. A. Stringham *et al.*, *Curr Biol.* **22**, 302 (2012).
13. D. H. Alexander, J. Novembre, K. Lange, *Genome Res.* **19**, 1655 (2009).
14. W. M. Levi, *The Pigeon* (Levi Publishing, Sumpter, SC, ed. 2 revised, 1986).
15. T. Amundsen, *Trends Ecol. Evol.* **15**, 149 (2000).
16. I. W. McKinnell, H. Makarenkova, I. de Curtis, M. Turmaine, K. Patel, *Dev. Biol.* **270**, 94 (2004).
17. R. N. Kelsh, M. L. Harris, S. Colanesi, C. A. Erickson, *Semin. Cell Dev. Biol.* **20**, 90 (2009).
18. M. Yandell *et al.*, *Genome Res.* **21**, 1529 (2011).
19. M. E. Elder *et al.*, *J. Immunol.* **166**, 656 (2001).
20. P. F. Colosimo *et al.*, *Science* **307**, 1928 (2005).
21. N. B. Sutter *et al.*, *Science* **316**, 112 (2007).
22. A. S. Van Laere *et al.*, *Nature* **425**, 832 (2003).
23. C. Mou *et al.*, *PLoS Biol.* **9**, e1001028 (2011).
24. Y. Wang *et al.*, *PLoS ONE* **7**, e34012 (2012).
25. L. F. Baptista, J. E. Gomez Martinez, H. M. Horblit, *Acta Zoologica Mex.* **25**, 719 (2009).

Acknowledgments: We thank the University of Washington Burke Museum for the *C. rupestris* tissue sample (UWBM 59803); J. Oldham, K. Wright, the Utah Pigeon Club, the National Pigeon Association, and A. and H. O. Christiansen for domestic pigeon samples; D. Clayton for feral samples; and D. Clayton, M. Horvath, D. Kingsley, R. Nielsen, and W. Warren for discussion and comments. Supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, NSF CAREER DEB-1149160, University of Utah Research Foundation (M.D.S.); NIH training grant T32GM007464 (S.A.S.); NIH training grant T32HD07491 (E.T.D.); an NSF EDEN internship (A.I.V.); NIH/NHGRI R01HG004694 and NIH ARRA GO RC2HG005619 (M.Y.); and the Danish National Research Foundation (M.T.P.G. and E.W.). We acknowledge a computer time allocation from the Center for High Performance Computing at the University of Utah. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession no. AKCR00000000 (the first version described here is AKCR01000000; raw reads, SRA052637); RNA-seq data for annotation, accession no. GSE39333; and raw reads for resequenced genomes, accession no. SRA054391.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1230422/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S27
Tables S1 to S28
References (*26*–*72*)

19 September 2012; accepted 7 December 2012
Published online 31 January 2013;
10.1126/science.1230422

KNOX2 Genes Regulate the Haploid-to-Diploid Morphological Transition in Land Plants

Keiko Sakakibara,[1,2,3,4]* Sayuri Ando,[3,4] Hoichong Karen Yip,[5]
Yosuke Tamada,[4,6] Yuji Hiwatashi,[4,6]† Takashi Murata,[4,6] Hironori Deguchi,[1]
Mitsuyasu Hasebe,[3,4,6] John L. Bowman[2,5]*

Unlike animals, land plants undergo an alternation of generations, producing multicellular bodies in both haploid (1n: gametophyte) and diploid (2n: sporophyte) generations. Plant body plans in each generation are regulated by distinct developmental programs initiated at either meiosis or fertilization, respectively. In mosses, the haploid gametophyte generation is dominant, whereas in vascular plants—including ferns, gymnosperms, and angiosperms—the diploid sporophyte generation is dominant. Deletion of the class 2 KNOTTED1-LIKE HOMEOBOX (KNOX2) transcription factors in the moss *Physcomitrella patens* results in the development of gametophyte bodies from diploid embryos without meiosis. Thus, KNOX2 acts to prevent the haploid-specific body plan from developing in the diploid plant body, indicating a critical role for the evolution of KNOX2 in establishing an alternation of generations in land plants.

Plants have a life cycle characterized by alternation between two generations, haploid (gametophyte) and diploid (sporophyte), where each phase develops a multicellular body (*1*, *2*). The gametophyte produces gametes—sperm (or pollen) and egg cells—and may be the dominant photosynthetic generation, as in liverworts, mosses, and hornworts. The sporophyte produces haploid spores via meiosis and is the dominant photosynthetic generation in the vascular plants. The alternation of generations in land plants results in the possibility that tissue differentiation in each generation is governed by different genetic programs, initiated by either fertilization (haploid to diploid) or meiosis (diploid to haploid). Land plants probably evolved from a freshwater algal ancestor, with a life cycle similar to

[1]Department of Biological Science, Graduate School of Science, Hiroshima University, Higashi-Hiroshima 739-8526, Japan. [2]School of Biological Sciences, Monash University, Melbourne, Victoria 3800, Australia. [3]ERATO, Japan Science and Technology Agency, Okazaki 444-8585, Japan. [4]National Institute for Basic Biology, Okazaki 444-8585, Japan. [5]Section of Plant Biology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA. [6]School of Life Science, The Graduate University for Advanced Studies, Okazaki 444-8585, Japan.

*To whom correspondence should be addressed. E-mail: john.bowman@monash.edu (J.L.B.); bara@hiroshima-u.ac.jp (K.S.)
†Present address: National Plant Phenomics Centre, Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3EB, UK.

CHAPTER 4

EPISTATIC AND COMBINATORIAL EFFECTS OF PIGMENTARY GENE MUTATIONS IN THE DOMESTIC PIGEON

Reprinted from Current Biology, 24, Eric T. Domyan, Michael W. Guernsey, Zev Kronenberg, Shreyas Krishnan, Raymond E. Boissy, Anna I. Vickrey, Clifford Rodgers, Pamela Cassidy, Sancy A. Leachman, John W. Fondon, Mark Yandell, Michael D. Shapiro , Epistatic and Combinatorial Effects of Pigmentary Gene Mutations in the Domestic Pigeon, 459-464, Copyright (2014), with permission from Elsevier Current Biology

4.1 Contribution

In this project, I had two roles. First, I called genetic variants for all the re-sequenced birds. This included re-phasing the variants and many quality assurance steps. My second role was assisting Eric Domyan with the GPA analyses. For the GPA studies we ran VAAST and GPAT++. The GPAT++ analyses (pFst) revealed an association between noncoding variants upstream of *Sox10* and the recessive red phenotype. The candidate genetic variant, the *e1* allele, is a ~7Kb that removes a melanocyte specific enhancer. The discovery of the *e1* allele encouraged me to start working on WHAM, for structural variant GPA.

Current Biology *24*, 459–464, February 17, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2014.01.020

Report

Epistatic and Combinatorial Effects of Pigmentary Gene Mutations in the Domestic Pigeon

Eric T. Domyan,[1] Michael W. Guernsey,[1] Zev Kronenberg,[2] Shreyas Krishnan,[3] Raymond E. Boissy,[4] Anna I. Vickrey,[1] Clifford Rodgers,[3] Pamela Cassidy,[5,6] Sancy A. Leachman,[5,6] John W. Fondon III,[3] Mark Yandell,[2] and Michael D. Shapiro[1,*]

[1]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA
[2]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA
[3]Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA
[4]Department of Dermatology, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA
[5]Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA
[6]Department of Dermatology, Oregon Health & Science University, Portland, OR 97239, USA

Summary

Understanding the molecular basis of phenotypic diversity is a critical challenge in biology, yet we know little about the mechanistic effects of different mutations and epistatic relationships among loci that contribute to complex traits. Pigmentation genetics offers a powerful model for identifying mutations underlying diversity and for determining how additional complexity emerges from interactions among loci. Centuries of artificial selection in domestic rock pigeons (*Columba livia*) have cultivated tremendous variation in plumage pigmentation through the combined effects of dozens of loci. The dominance and epistatic hierarchies of key loci governing this diversity are known through classical genetic studies [1–6], but their molecular identities and the mechanisms of their genetic interactions remain unknown. Here we identify protein-coding and *cis*-regulatory mutations in *Tyrp1*, *Sox10*, and *Slc45a2* that underlie classical color phenotypes of pigeons and present a mechanistic explanation of their dominance and epistatic relationships. We also find unanticipated allelic heterogeneity at *Tyrp1* and *Sox10*, indicating that color variants evolved repeatedly though mutations in the same genes. These results demonstrate how a spectrum of coding and regulatory mutations in a small number of genes can interact to generate substantial phenotypic diversity in a classic Darwinian model of evolution [7].

Results and Discussion

In the domestic rock pigeon (*Columba livia*), hundreds of years of accumulated experience by amateur and professional geneticists provide strong evidence that many complex color traits can be partitioned into combined effects of multiple loci, and that the same loci control similar traits across breeds [6]. The classical major color locus (*B*) is a sex-linked gene that confers one of three "base" colors [1–5]: wild-type blue/black (B^+), ash-red (B^A), and brown (*b*) (Figures 1A–1C). The B^A allele is dominant to B^+ and *b*, and *b* is recessive to the others. Blue/black and brown phenotypes result from high amounts of eumelanin and low amounts of pheomelanin; melanin ratios are reversed in ash-red birds [8]. In addition, the autosomal recessive mutation *recessive red* (*e*) acts epistatically to the *B* locus to elevate pheomelanin production, generating red plumage color irrespective of *B* locus genotype [2, 8] (Figure 1D). Mutant alleles of a third locus, the sex-linked recessive *dilute* (*d*), interact additively with *B* and *e* to lighten plumage color and further enrich pigmentation diversity [1, 2, 8] (Figures 1E–1H). This detailed Mendelian understanding of key phenotypes provides a robust foundation to investigate how genes and alleles interact to generate color variation. However, the molecular basis of this diversity—including the identities of genes underlying major pigmentation variants and a mechanistic explanation for their intra- and interlocus interactions—remains unknown [9, 10].

Multiple Mutations in *Tyrp1* Underlie Base Color Variation in Pigeons

Previously, we reported whole-genome sequences for 41 rock pigeons [11] with diverse color phenotypes. To investigate the molecular identity of the *B* color locus, we compared the genomes of 6 ash-red pigeons to 26 blue/black pigeons for coding changes associated with pigmentation phenotypes using the Variant Annotation, Analysis, and Search Tool (VAAST) [12]. A single gene achieved genome-wide significance: tyrosinase-related protein 1 (*Tyrp1*) ($p = 1.3 \times 10^{-6}$; see Figure S1A available online), which encodes a key enzyme in the melanin synthesis pathway. All blue/black pigeons were homozygous G on the *Tyrp1* sense strand at position 214991 on genomic scaffold 6 (B^+ allele), whereas ash-red pigeons were hetero- or homozygous for C (B^A allele), consistent with the dominant mode of inheritance of ash-red. The B^A mutation causes an alanine-to-proline substitution at codon 23 (A23P), corresponding to the cleavage site of the signal peptide (Figure 2A). In addition to finding a single haplotype containing the B^A allele in our whole-genome panel (Figure S1B), we found a perfect association between the dominant B^A mutation and the ash-red phenotype in an additional 49 ash-red birds from 20 breeds, and 105 blue/black or brown birds from 36 breeds (Figure 2B). These results suggest that the ash-red mutation occurred only once and spread species-wide through selective breeding, similar to our previous finding that the same mutation in *EphB2* underlies the head crest phenotype in multiple pigeon breeds [11].

Quantitative RT-PCR analysis revealed that *Tyrp1* mRNA levels from developing feathers of B^+ and B^A pigeons were indistinguishable (Figure S1C); however, the location of the B^A mutation at the highly conserved cleavage site of the signal peptide (Figure S1E) suggested that cleavage efficiency might be affected. We therefore expressed N- and C-terminally tagged B^+ and B^A versions of TYRP1 protein in cell culture, and we found that cleavage efficiency was dramatically reduced by the B^A mutation (relative efficiency: $B^+ = 1 \pm 0.18$, $B^A = 0.14 \pm 0.04$; n = 4 independent transfections each;

*Correspondence: shapiro@biology.utah.edu

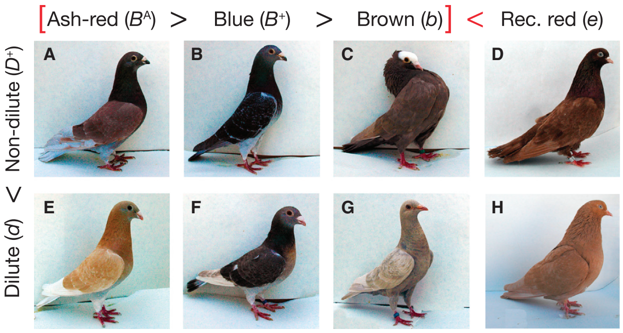

[Ash-red (B^A) > Blue (B^+) > Brown (b)] < Rec. red (e)
Non-dilute (D^+)
>
Dilute (d)
A
B
C
D
E
F
G
H

Figure 1. Common Color Phenotypes of Domestic Rock Pigeons

(A–C) Allelic variation at sex-linked major color locus (*B*).

(D) Recessive autosomal mutation *recessive red* (*e*).

(E–H) Recessive allele at another sex-linked locus, *dilute* (*d*), reduces color intensity to generate ash-yellow (E), dun (F), khaki (G), or recessive yellow (H) phenotypes.

Black chevrons (>) indicate order of dominance among alleles; red chevron indicates epistasis between loci. Breeds shown are show-type racing homer (A, B, and E–G), mookee (C), parlor roller (D), and Birmingham roller (H).

$p < 0.002$) (Figure 2C). Furthermore, spatial organization of pigment synthesis differed between *B* genotypes: premelanosomes in regenerating blue/black (B^+) feathers had a well-organized, lamellar matrix, and melanosomes were darkly pigmented, whereas ash-red (B^A and B^+B^A) feathers had a disorganized matrix and only lightly pigmented melanosomes (Figure 2D). After incubation with the melanin precursor L-DOPA, melanosomes from both wild-type and ash-red birds became darkly pigmented, indicating normal catalytic activity of the melanogenic enzyme tyrosinase (TYR) in ash-red birds. However, pigment synthesis in B^+ feathers showed strongest staining localized to the limiting membrane of the melanosome (Figure 2D), whereas staining was diffuse in melanosomes from B^A and B^+B^A feathers. Thus, the striking reduction in TYRP1 cleavage efficiency may disrupt the spatial organization of pigment synthesis activity, providing insight into the molecular basis of dominance of the B^A allele. The dominant *Light* (B^{lt}) *Tyrp1* allele of mice, a missense mutation near the same cleavage site, causes melanocyte death, probably through the accumulation of cytotoxic pigment intermediates [13]. Whether a similar accumulation of cytotoxins contributes to the pheomelanic phenotype of ash-red pigeons is unknown. However, unlike the mouse B^{lt} allele, the pigeon B^A allele results in a different kind and localization of melanin production rather than abrogation of melanogenesis.

In contrast to the single ash-red mutation, *Tyrp1* sequences from 51 brown pigeons from 30 breeds revealed three different nonsense and frameshift mutations (R72X, b^1;*411-418 del*, b^2 ;*893 delA*, b^3) (Figures 2A and S1D), predicted to be null alleles. Indeed, *Tyrp1* mRNA abundance in b^3 pigeons—the most common *b* allele in our sample—is greatly reduced or absent (relative expression: $B^+ = 1 \pm 0.53$, $n = 4$; $b^3 = 0.009 \pm 0.005$, $n = 3$; $p = 0.05$) (Figure 2E), consistent with the activation of nonsense-mediated decay. This indicates that, in contrast to the single origin of the ash-red phenotype, brown color has evolved multiple times in pigeons. Several brown pigeons did not have any of the identified *b* alleles, raising the possibility that additional mutations might also cause brown feather color (Figure S1D).

Together, these results demonstrate distinct effects of different mutations in the same gene and also confirm the predicted orthology of the classical mouse and pigeon *B* loci [5, 14]. Our analyses suggest a model in which B^A is a neomorphic allele that alters processing of the mutant TYRP1 protein within the cell. Since TYRP1 can modulate TYR activity [15–18], we postulate that the B^A version of TYRP1 protein alters normal TYR functionality, resulting in an increased ratio of pheomelanin to eumelanin production. In contrast, *Tyrp1* loss-of-function alleles b^1–b^3 cause brown pigment production, consistent with findings in other vertebrates [19].

Recurrent Deletions of a *Sox10* Enhancer Underlie Recessive Red

In addition to the dominant, sex-linked B^A allele, the autosomal mutation *recessive red* (*e*) acts epistatically to *B* to increase pheomelanogenesis and decrease eumelanogenesis (Figure 1D). VAAST scans for the *e* locus did not identify a strong candidate, suggesting that an unannotated structural variant, such as a large insertion or deletion, might underlie this phenotype. To identify candidates for *e*, we compared expression of several genes involved in melanin biosynthesis and found that the transcription factor *Sox10* and one of its target genes, *Tyrp1* (the *B* locus), were significantly downregulated in feathers of recessive red birds (Figure 3A) (*Sox10* relative expression: blue/black = 1 ± 0.62; recessive red = 0.14 ± 0.07, $p = 0.001$; *Tyrp1* relative expression: blue/black = 1 ± 0.556, recessive red = 0.0001 ± 0.00006, $p = 0.002$; $n = 6$ for each). Other melanin biosynthesis genes did not show altered transcript levels (Figure S2A), suggesting that a mutation directly or indirectly affecting *Sox10* expression might underlie the recessive red phenotype.

Deletions of a conserved *Sox10* enhancer result in pigmentation defects in other vertebrates, including a lack of pigmentation in mouse and increased pheomelanin production in chicken [20, 21]. Alignment of the pigeon reference genome assembly (a recessive red Danish tumbler [11]) upstream of *Sox10* to the orthologous regions of the chicken and zebra finch genomes identified a 7.5 kb deletion in the pigeon genome (Figures 3B and S2B). Furthermore, four recessive red birds in our genome resequencing panel—but no wild-type birds—were homozygous for this deletion. Importantly, the deletions in pigeon, chicken, and mouse all span a conserved enhancer element that drives *Sox10* expression in melanocytes [20, 23].

To test for broader association between the pigeon *Sox10* enhancer deletion and recessive red, we genotyped 41 recessive red pigeons from 19 breeds and 103 wild-type pigeons from 45 breeds. We found that 21 recessive red birds (but no wild-type birds) were homozygous for the deletion harbored by the reference genome (e^1 allele; Figures 3B and 3C). An additional 17 of the recessive red birds (but no

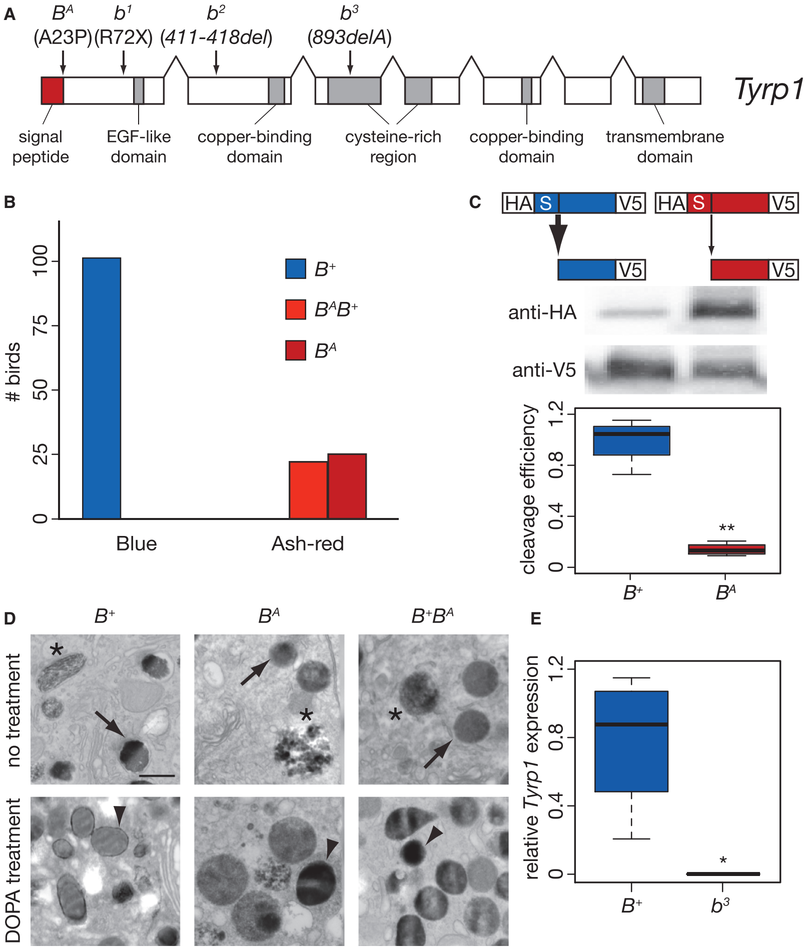


Figure 2. *Tyrp1* Is the Major Color Locus *B* in Domestic Pigeons

(A) Schematic of the genomic *Tyrp1* locus with putative B^A and *b* mutations.

(B) Histogram of genotypes of pigeons displaying wild-type or ash-red phenotypes.

(C) Schematic and western blot analysis of cleavage of TYRP1 proteins encoded by B^+ and B^A alleles, demonstrating reduced cleavage efficiency of the B^A allele. HA, N-terminal hemagglutinin epitope tag; V5, C-terminal V5 epitope tag. Boxes in (C) and (E) span first to third quartiles; bars extend to minimum and maximum observed values; black line indicates median. $^{**}p < 0.002$.

(D) Ultrastructural analysis of melanocytes from B^+, B^A, and B^+B^A feathers. Asterisks indicate premelanosomes, arrows indicate untreated melanosomes, and arrowheads indicate DOPA-treated melanosomes. Scale bar represents 500 nm.

(E) *Tyrp1* mRNA abundance in B^+ and b^3 feathers by qRT-PCR. $^{*}p = 0.05$.

wild-type birds) were homozygous for a second, 2.5 kb deletion (e^2) that partially overlaps e^1, and the remaining three birds were heterozygous e^1e^2 (Figures 3B and 3C). Since both pigeon deletions span the *Sox10* melanocyte enhancer, we predicted that the reduction in *Sox10* expression in recessive red birds was due to a *cis*-regulatory change. We therefore assayed allele-specific expression of E^+ and e^2 alleles of *Sox10* in E^+e^2 heterozygous birds, and we found that the e^2 allele was expressed at only ~10% of E^+ levels (SNP1 = 0.126 ± 0.055, SNP2 = 0.056 ± 0.043, SNP3 = 0.127 ± 0.059; $p < 0.0001$ for each, n = 10 E^+e^2 birds) (Figure 3D). Since in heterozygotes both the E^+ and e^2 alleles

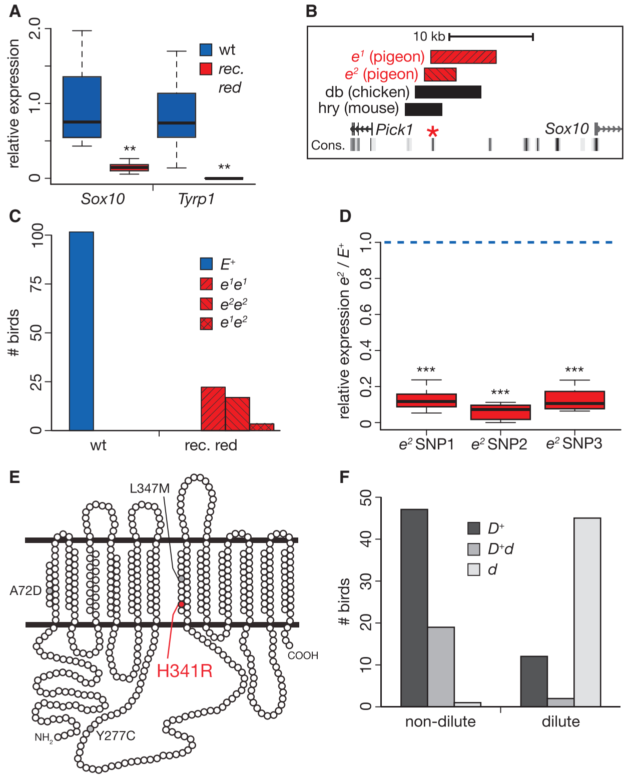


Figure 3. *Sox10* and *Slc45a2* Are the *recessive red* (*e*) and *dilute* (*d*) Loci of Domestic Pigeons

(A) qRT-PCR analysis of *Sox10* and *Tyrp1* in wild-type versus recessive red feathers (see Figure S2 for additional genes). Boxes in (A) and (D) span first to third quartiles; bars extend to minimum and maximum observed values; black line indicates median. $^{**}p \leq 0.002$.

(B) Schematic of deletions upstream of *Sox10* in recessive red pigeons (e^1, e^2), dark-brown (db) chicken [20], and *Hry* mutant mouse [21]. Red asterisk denotes a conserved element deleted in all three species. Conservation track is based on Multiz alignment to chicken, human, mouse, rat, opossum, *Xenopus tropicalis*, and zebrafish in UCSC Genome Browser (http://genome.ucsc.edu/; chicken assembly v2.1 used as framework).

(C) Histogram of genotypes of pigeons displaying wild-type or recessive red phenotypes.

(D) Expression of SNPs in the e^2 allele relative to the E^+ allele of *Sox10* in feathers of E^+e^2 heterozygous pigeons. Blue dashed line indicates normalized expression level of E^+ allele. $^{***}p < 0.0001$.

(E) Schematic of SLC45A2 protein with putative *d* mutation in red. Mutations in chicken and quail associated with lightened feather color are indicated in gray. Adapted from [22].

(F) Histogram of genotypes of pigeons displaying wild-type or dilute phenotypes.

are in the same cellular environment, this experiment confirmed that the reduction in *Sox10* expression from the e^2 allele is due to a *cis*-acting mutation. Together, these genetic and expression results implicate the deletion of a *Sox10* melanocyte enhancer as the molecular basis of recessive red in domestic pigeons (Figure 3B). These results also demonstrate that the *E* (*extension*) loci of mammals (*Mc1r*) and pigeons (*Sox10*) are not orthologous [5, 9, 24–26]. Moreover, similar to the brown phenotype, recessive red appears to have evolved more than once in pigeons. While we do not observe obvious phenotypic distinctions between e^1 and e^2 homozygotes, it is possible that the different deletions generate subtly different effects on color by altering other unidentified regulatory elements [27].

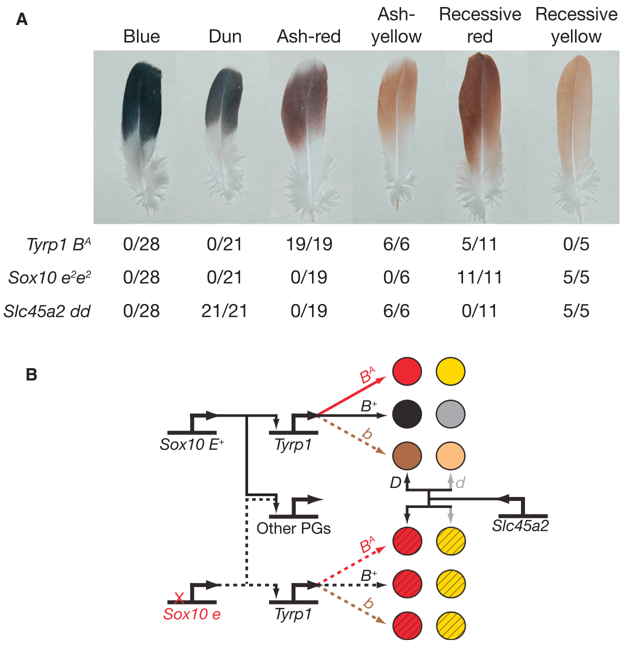
The epistatic relationship of *e* to *B* is now easily reconciled in light of their molecular identities and mutations: *Sox10* directly regulates *Tyrp1* expression in melanocytes [28] (Figure 3A), explaining how loss of *Sox10* expression abrogates phenotypic effects of *Tyrp1* genotypes. Interestingly, the recessive red phenotype caused by *Sox10* downregulation is distinct from the brown phenotype of *Tyrp1* loss-of-function mutants, possibly owing to contributions of additional *Sox10* regulatory targets or residual *b* allele activity.

Missense Mutation in *Slc45a2* Is Associated with Color Dilution

While the *B* and *E* loci affect pigment color, the sex-linked recessive *dilute* (*d*) reduces pigment quantity, further enriching pigmentation diversity [8] (Figures 1E–1H). To identify candidates for *d*, we compared the genomes of 5 birds with diluted feather color and 31 birds with nondiluted pigment intensity using VAAST. A single gene achieved genome-wide significance: solute carrier family 45 member 2 (*Slc45a2*, $p = 2.65 \times 10^{-6}$; Figure S3A), which is associated with pigmentation phenotypes in diverse vertebrates, including other birds [22, 29–32], but is not orthologous to the *dilute* locus in mouse (*Myo5a*) [33]. In pigeons, the *d* mutation causes a histidine-to-arginine substitution (H341R) at a highly conserved intramembrane residue of SLC45A2 (Figures 3E and S3C). We genotyped an additional 59 diluted birds from 26 breeds and 67 nondiluted birds from 41 breeds and found a strong (but not perfect) association between *d* genotypes and color intensity under a recessive model (Fisher's exact test, $p < 2.2 \times 10^{-16}$) (Figure 3F). Fourteen birds not homozygous for *d* had diluted feather color, and one homozygote was reported to have nondiluted color. However, several other loci can cause either lightened (e.g., *milky*, *reduced*, and *faded*) or darkened (e.g., *dirty*, *sooty*, and *smoky*) pigmentation in pigeons [6], and it is expected that a broad hobbyist-identified sample should include birds with varied genetic bases for color intensity.

Mutations and Color Traits Cosegregate in a Controlled Cross

As an independent test of our association analyses, we examined cosegregation of pigmentation phenotypes and our three



	Blue	Dun	Ash-red	Ash-yellow	Recessive red	Recessive yellow
Tyrp1 B^A	0/28	0/21	19/19	6/6	5/11	0/5
Sox10 e^2e^2	0/28	0/21	0/19	0/6	11/11	5/5
Slc45a2 dd	0/28	21/21	0/19	6/6	0/11	5/5

Figure 4. Segregation Analysis and Mechanistic Model of Common Color Phenotypes

(A) Representative feathers from F_1 and F_2 offspring in a cross segregating B^A, e^2, and *d* mutations; numbers of birds with a given genotype are listed below each phenotype.

(B) Schematic illustrating common feather colors in pigeons and mutations responsible for their production. Other pigmentation genes (PGs) are probably also affected by a decrease in *Sox10* expression, thereby causing differences between *b* (*Tyrp1* loss of function) and *e* (*Sox10* loss of function) phenotypes.

candidate loci in progeny of a male pigeon doubly heterozygous for B^A and *d* alleles (B^AB^+,D^+d,E^+E^+) mated to two recessive yellow females (B^+,*d*,e^2e^2; females have heterogametic sex chromosomes and are therefore hemizygous at the sex-linked *B* and *D* loci). As anticipated, segregation of the ash-red (dominant) and dilute phenotypes was observed in the F_1 and F_2 generations, whereas recessive red was observed only in the F_2. We genotyped all three generations for the candidate mutations in *Tyrp1*, *Sox10*, and *Slc45a2* and found complete cosegregation between alleles at these loci and their respective pigmentation phenotypes (Figure 4A). Coupled with our genetic association results, these transmission genetics data strongly support the molecular identities of the classical B^A (and *b*), *e*, and *d* alleles as mutations in *Tyrp1*, *Sox10*, and *Slc45a2*, respectively.

Few Loci Generate Many Phenotypes

Similar to the genetic architecture of dog coat morphology [34], a relatively small number of loci generate a wide range of plumage color phenotypes in pigeons (Figure 4B). We found that coding changes at the base color locus *B* result in a neomorphic dominant allele (B^A, ash-red) that interferes with melanosome formation and localization of melanogenesis to cause one derived phenotype, and multiple recessive, putative null alleles (b^1–b^3, brown) that underlie another phenotype. The *e* mutation is epistatic to *B* genotypes due to a regulatory mutation in *Sox10*, which is a transcriptional regulator of *Tyrp1*. A mutation at the *d* locus influences the color of birds of all genotypes at *B* and *e* by reducing the quantity of pigment produced, generating an additional layer of genetic complexity and phenotypic diversity. We find evidence that some color phenotypes, such as ash-red, appear to have a single origin whereas others, such as brown and recessive red, originated multiple times.

Many color phenotypes in pigeons and other domestic animals result from artificial selection [35]. Nevertheless, the combination of coding and regulatory changes, combined effects of multiple loci on a common phenotypic output, and single and multiple origins of derived alleles is reminiscent of the genetic architecture of a variety of adaptive traits in the wild (e.g., [27, 36–38]). Additionally, the specific genes that we have implicated in plumage color phenotypes in pigeons also contribute to both natural pigmentation diversity and skin disease in humans, including melanoma risk [39, 40]. Thus, by elucidating the complex interactions among these loci, we enrich our mechanistic understanding of adaptive and nonadaptive variation across species.

Accession Numbers

Sequence data reported in this paper have been deposited in GenBank with the accession numbers KJ023250–KJ023253.

Supplemental Information

Supplemental Information includes three figures, Supplemental Experimental Procedures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.01.020.

Acknowledgments

All animal protocols were approved by the Institutional Animal Care and Use Committees of the University of Utah (protocols 09-04015 and 10-05007) and the University of Texas at Arlington (protocol 09-009). We thank members of the Shapiro and Fondon labs for sample collection assistance. This work was supported by NIH fellowships T32HD07491 and F32GM103077 (E.T.D.); NIH training grant T32GM007464 (Z.K.); an NSF EDEN internship (A.I.V.); Huntsman Cancer Foundation and the Tom C. Mathews Jr. Familial Melanoma Research Clinic Endowment (S.A.L.); NIH grants R01HG004694, R01GM104390, and ARRA GO RC2HG005619 (M.Y.); and a Burroughs Wellcome Fund Career Award in the Biomedical Sciences and NSF CAREER DEB1149160 (M.D.S.). We acknowledge computer time allocation from the University of Utah Center for High Performance Computing, Trisha Wittkopp for technical advice, and Dale Clayton for husbandry assistance. We gratefully acknowledge years of mentoring and assistance by master breeder Brad R. Child, and we dedicate this work to his memory.

Received: November 6, 2013
Revised: January 7, 2014
Accepted: January 9, 2014
Published: February 6, 2014

References

1. Cole, L.J. (1912). A case of sex-linked inheritance in the domestic pigeon. Science *36*, 190–192.
2. Cole, L.J., and Kelley, F.J. (1919). Studies on inheritance in pigeons. III. Description and linkage relations of two sex-linked characters. Genetics *4*, 183–203.
3. Christie, W., and Wriedt, C. (1927). Schokolade, ein neuer geschlechtsgebundener Farbencharakter bei Tauben. Z. Indukt. Abstamm. Vererbungsl. *43*, 391–392.
4. Hawkins, L.E. (1931). Studies on inheritance in pigeons. X. Relation of chocolate to black and dominant red. Genetics *16*, 547–573.
5. Steele, D.G. (1931). Studies on inheritance in pigeons. IX. The chocolate-brown plumage color. Genetics *16*, 532–546.
6. Sell, A. (1994). Breeding and Inheritance in Pigeons (Hengersberg: Schober Verlags-GmbH).
7. Darwin, C. (1859). On the Origin of Species by Means of Natural Selection (London: John Murray).
8. Haase, E., Ito, S., Sell, A., and Wakamatsu, K. (1992). Melanin concentrations in feathers from wild and domestic pigeons. J. Hered. *83*, 64–67.
9. Guernsey, M.W., Ritscher, L., Miller, M.A., Smith, D.A., Schöneberg, T., and Shapiro, M.D. (2013). A Val85Met mutation in melanocortin-1 receptor is associated with reductions in eumelanic pigmentation and cell surface expression in domestic rock pigeons (*Columba livia*). PLoS ONE *8*, e74475.
10. Derelle, R., Kondrashov, F.A., Arkhipov, V.Y., Corbel, H., Frantz, A., Gasparini, J., Jacquin, L., Jacob, G., Thibault, S., and Baudry, E. (2013). Color differences among feral pigeons (*Columba livia*) are not attributable to sequence variation in the coding region of the melanocortin-1 receptor gene (MC1R). BMC Res. Notes *6*, 310.
11. Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., et al. (2013). Genomic diversity and evolution of the head crest in the rock pigeon. Science *339*, 1063–1067.
12. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.
13. Johnson, R., and Jackson, I.J. (1992). Light is a dominant mouse mutation resulting in premature cell death. Nat. Genet. *1*, 226–229.
14. Zdarsky, E., Favor, J., and Jackson, I.J. (1990). The molecular basis of brown, an old mouse mutation, and of an induced revertant to wild type. Genetics *126*, 443–449.
15. Manga, P., Sato, K., Ye, L., Beermann, F., Lamoreux, M.L., and Orlow, S.J. (2000). Mutational analysis of the modulation of tyrosinase by tyrosinase-related proteins 1 and 2 in vitro. Pigment Cell Res. *13*, 364–374.
16. Jiménez-Cervantes, C., Martínez-Esparza, M., Solano, F., Lozano, J.A., and García-Borrón, J.C. (1998). Molecular interactions within the melanogenic complex: formation of heterodimers of tyrosinase and TRP1 from B16 mouse melanoma. Biochem. Biophys. Res. Commun. *253*, 761–767.
17. Hearing, V.J., Tsukamoto, K., Urabe, K., Kameyama, K., Montague, P.M., and Jackson, I.J. (1992). Functional properties of cloned melanogenic proteins. Pigment Cell Res. *5*, 264–270.
18. Kobayashi, T., and Hearing, V.J. (2007). Direct interaction of tyrosinase with Tyrp1 to form heterodimeric complexes in vivo. J. Cell Sci. *120*, 4261–4268.
19. Hubbard, J.K., Uy, J.A., Hauber, M.E., Hoekstra, H.E., and Safran, R.J. (2010). Vertebrate pigmentation: from underlying genes to adaptive function. Trends Genet. *26*, 231–239.
20. Gunnarsson, U., Kerje, S., Bed’hom, B., Sahlqvist, A.S., Ekwall, O., Tixier-Boichard, M., Kämpe, O., and Andersson, L. (2011). The dark brown plumage color in chickens is caused by an 8.3-kb deletion upstream of SOX10. Pigment Cell Melanoma Res. *24*, 268–274.
21. Antonellis, A., Bennett, W.R., Menheniott, T.R., Prasad, A.B., Lee-Lin, S.Q., Green, E.D., Paisley, D., Kelsh, R.N., Pavan, W.J., and Ward, A.; NISC Comparative Sequencing Program (2006). Deletion of long-range sequences at Sox10 compromises developmental expression in a mouse model of Waardenburg-Shah (WS4) syndrome. Hum. Mol. Genet. *15*, 259–271.
22. Gunnarsson, U., Hellström, A.R., Tixier-Boichard, M., Minvielle, F., Bed’hom, B., Ito, S., Jensen, P., Rattink, A., Vereijken, A., and Andersson, L. (2007). Mutations in SLC45A2 cause plumage color variation in chicken and Japanese quail. Genetics *175*, 867–877.
23. Antonellis, A., Huynh, J.L., Lee-Lin, S.Q., Vinton, R.M., Renaud, G., Loftus, S.K., Elliot, G., Wolfsberg, T.G., Green, E.D., McCallion, A.S., and Pavan, W.J. (2008). Identification of neural crest and glial enhancers at the mouse Sox10 locus through transgenesis in zebrafish. PLoS Genet. *4*, e1000174.
24. Joerg, H., Fries, H.R., Meijerink, E., and Stranzinger, G.F. (1996). Red coat color in Holstein cattle is associated with a deletion in the MSHR gene. Mamm. Genome *7*, 317–318.
25. Newton, J.M., Wilkie, A.L., He, L., Jordan, S.A., Metallinos, D.L., Holmes, N.G., Jackson, I.J., and Barsh, G.S. (2000). Melanocortin 1 receptor variation in the domestic dog. Mamm. Genome *11*, 24–30.
26. Robbins, L.S., Nadeau, J.H., Johnson, K.R., Kelly, M.A., Roselli-Rehfuss, L., Baack, E., Mountjoy, K.G., and Cone, R.D. (1993). Pigmentation phenotypes of variant extension locus alleles result from point mutations that alter MSH receptor function. Cell *72*, 827–834.
27. Linnen, C.R., Poh, Y.P., Peterson, B.K., Barrett, R.D., Larson, J.G., Jensen, J.D., and Hoekstra, H.E. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. Science *339*, 1312–1316.
28. Murisier, F., Guichard, S., and Beermann, F. (2006). A conserved transcriptional enhancer that specifies Tyrp1 expression to melanocytes. Dev. Biol. *298*, 644–655.
29. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
30. Mariat, D., Taourit, S., and Guérin, G. (2003). A mutation in the MATP gene causes the cream coat colour in the horse. Genet. Sel. Evol. *35*, 119–133.
31. Xu, X., Dong, G.X., Hu, X.S., Miao, L., Zhang, X.L., Zhang, D.L., Yang, H.D., Zhang, T.Y., Zou, Z.T., Zhang, T.T., et al. (2013). The genetic basis of white tigers. Curr. Biol. *23*, 1031–1035.
32. Fukamachi, S., Shimada, A., and Shima, A. (2001). Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. Nat. Genet. *28*, 381–385.
33. Mercer, J.A., Seperack, P.K., Strobel, M.C., Copeland, N.G., and Jenkins, N.A. (1991). Novel myosin heavy chain encoded by murine dilute coat colour locus. Nature *349*, 709–713.
34. Cadieu, E., Neff, M.W., Quignon, P., Walsh, K., Chase, K., Parker, H.G., Vonholdt, B.M., Rhue, A., Boyko, A., Byers, A., et al. (2009). Coat variation in the domestic dog is governed by variants in three genes. Science *326*, 150–153.
35. Andersson, L. (2009). Studying phenotypic evolution in domestic animals: a walk in the footsteps of Charles Darwin. Cold Spring Harb. Symp. Quant. Biol. *74*, 319–325.
36. Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Jr., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science *327*, 302–305.
37. Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G., Jr., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D., and Kingsley, D.M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science *307*, 1928–1933.
38. Steiner, C.C., Weber, J.N., and Hoekstra, H.E. (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol. *5*, e219.
39. Jablonski, N.G. (2012). Human skin pigmentation as an example of adaptive evolution. Proc. Am. Philos. Soc. *156*, 45–57.
40. Law, M.H., Macgregor, S., and Hayward, N.K. (2012). Melanoma genetics: recent findings take us beyond well-traveled pathways. J. Invest. Dermatol. *132*, 1763–1774.

CHAPTER 5

DISAP-DEPENDENT STRIATED FIBER ELONGATION IS REQUIRED TO ORGANIZE CILIARY ARRAYS

©Galati et al., 2014. Originally published in *The Journal of Cell Biology*. doi: 10.1083/jcb.201409123.

5.1 Contribution

I mapped the disA-1 recessive allele to the *DisA* locus. While this sounds trivial, the *T. thermophila* genome is divided between two nuclei with similar DNA content. The duplicate genome makes short-read mapping very difficult. Other researchers were unable to map the disA-1 allele with the same data. I set up a special workflow for mapping mutations in *T. thermophila*. Unlike previous efforts, I expended the search for the disA-1 allele outside of coding sequences. The disA-1 allele is a splice acceptor mutation, which results in a truncated protein. This analysis is the first example of mapping a *T. thermophila* mutant using high throughput sequencing data. My contribution can be seen in supplemental Figure 1C-D.

Published December 22, 2014

DisAp-dependent striated fiber elongation is required to organize ciliary arrays

Domenico F. Galati,[1] Stephanie Bonney,[1] Zev Kronenberg,[2] Christina Clarissa,[3] Mark Yandell,[2] Nels C. Elde,[2] Maria Jerka-Dziadosz,[4] Thomas H. Giddings,[3] Joseph Frankel,[5] and Chad G. Pearson[1]

[1]Anschutz Medical Campus, Department of Cell and Developmental Biology, University of Colorado, Aurora, CO 80045
[2]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112
[3]Molecular, Cellular and Developmental Biology, University of Colorado at Boulder, Boulder, CO 80309
[4]Department of Cell Biology, M. Nencki Institute of Experimental Biology, 02-093 Warsaw, Poland
[5]Department of Biological Sciences, University of Iowa, Iowa City, IA 52242

Cilia-organizing basal bodies (BBs) are microtubule scaffolds that are visibly asymmetrical because they have attached auxiliary structures, such as striated fibers. In multiciliated cells, BB orientation aligns to ensure coherent ciliary beating, but the mechanisms that maintain BB orientation are unclear. For the first time in *Tetrahymena thermophila*, we use comparative whole-genome sequencing to identify the mutation in the BB disorientation mutant *disA-1*. disA-1 abolishes the localization of the novel protein DisAp to *T. thermophila* striated fibers (kinetodesmal fibers; KFs), which is consistent with DisAp's similarity to the striated fiber protein SF-assemblin. We demonstrate that DisAp is required for KFs to elongate and to resist BB disorientation in response to ciliary forces. Newly formed BBs move along KFs as they approach their cortical attachment sites. However, because they contain short KFs that are rotated, BBs in disA-1 cells display aberrant spacing and disorientation. Therefore, DisAp is a novel KF component that is essential for force-dependent KF elongation and BB orientation in multiciliary arrays.

Introduction

Motile cilia are whiplike projections that generate hydrodynamic force. Cilia-generated fluid flow is required for symmetry breaking during embryogenesis, mucus clearance, cerebrospinal fluid flow, and the directed movement of unicellular organisms (Marshall and Kintner, 2008). One cycle of ciliary beating constitutes a power stroke and a subsequent recovery stroke. Thus, ciliary beating is directional, and to produce coherent fluid flow, multiple cilia must orient their beating along a common plane, which is typically the cell's anterior–posterior axis. The importance of proper cilia orientation is underscored by the observation that cilia orientation defects accompany primary cilia dyskinesias, a devastating class of genetic disorders (Rayner et al., 1996).

Cilia are organized by cylindrical microtubule scaffolds called basal bodies (BBs) that dock at the cell cortex (Jana et al., 2014). BBs are innately asymmetric and their polarity is reflected in the attachment of auxiliary structures, including striated fibers (Allen, 1969; Pearson, 2014). Thus, BBs have a specific orientation that determines the direction of ciliary beating (Tamm et al., 1975; Gibbons, 1981; Hoops et al., 1984). BBs with improper orientation relative to the cellular anterior–posterior axis will disrupt cilia-generated fluid flow. The mechanisms that organize and maintain BB orientation remain ill-defined.

Striated fibers project asymmetrically from BBs and influence BB positioning by an unknown mechanism (Allen, 1967; Wright et al., 1983; Hoops et al., 1984). SF-assemblin and rootletin are coiled-coil proteins that self-organize into filamentous fiber structures and constitute major structural components of striated fibers in protists and vertebrates, respectively (Lechtreck and Melkonian, 1991; Yang et al., 2002), although other proteins are also present (Lechtreck and Melkonian, 1998; Park et al., 2008; Chien et al., 2013). Moreover, striated fibers display dynamic assembly and disassembly (Salisbury et al., 1984; Sperling et al., 1991; Francia et al., 2012). Thus, striated fibers are complex

Correspondence to Chad G. Pearson: Chad.Pearson@ucdenver.edu

Abbreviations used in this paper: BB, basal body; KF, kinetodesmal fiber; PEO, polyethylene oxide; TEM, transmission electron microscopy; WT, wild type.

Supplemental Material can be found at:
http://jcb.rupress.org/content/suppl/2014/12/19/jcb.201409123.DC1.html

The Rockefeller University Press $30.00
J. Cell Biol. Vol. 207 No. 6 705–715
www.jcb.org/cgi/doi/10.1083/jcb.201409123

Published December 22, 2014

and dynamic structures of which our molecular understanding is limited.

Unicellular ciliates, such as *Tetrahymena thermophila*, and multiciliated vertebrate cells harbor hundreds of cilia organized into ciliary arrays. Ciliary array BBs exhibit evolutionarily conserved striated fiber placement directly opposite the cilium's power stroke (Allen, 1969; Peraldi-Roux et al., 1991; Frankel, 1999). The *Tetrahymena* striated fiber, the kinetodesmal fiber (KF), emanates close to the BB's base and terminates within or directly underneath the membrane-skeletal layer near the adjacent anterior BB (Allen, 1967, 1969). The apposition of the KF and the postciliary microtubules from the anterior BB supports speculation that KFs stabilize ciliary rows by providing a physical linkage between neighboring ciliary units and by linkage to subcortical structures (Allen, 1967; Iftode and Fleury-Aubusson, 2003; Wloga and Frankel, 2012). Although this hypothesis has been strengthened by observations in *Chlamydomonas reinhardtii* (Wright et al., 1983; Hoops et al., 1984), a mechanistic understanding of how striated fibers organize ciliary arrays and respond to and resist mechanical forces has not been established.

Results and discussion

DisAp localizes to KFs and orients BBs

disA-1 is a single-locus, recessive mutation generated in a mutagenesis screen for *T. thermophila* BB organization defects (Frankel, 1979, 2008; Jerka-Dziadosz et al., 1995). *DISA* organizes BBs into ciliary rows, but is dispensable for global cellular polarity and for ciliogenesis (Fig. 1 A; Jerka-Dziadosz et al., 1995). The *disA-1* gene was identified using comparative genome sequence analysis with next-generation sequencing (Fig. S1, A and C). This approach identified a splice acceptor site mutation in intron 1 of a novel gene (TTHERM_00941400), which results in a severely truncated protein (Fig. 1 B and Fig. S1 B). The gene encodes a protein (DisAp) containing a similarity to the SF-assemblin consensus domain (Fig. 1 C; Lechtreck and Melkonian, 1998). Although the faint resemblance to SF-assemblin alerted us to a potential role in KF structure (pfam06705; BLASTp query of the Conserved Domain Database), phylogenetic analysis revealed that DisAp is a member of a distinct family of proteins conserved among ciliates, with seven paralogues in *T. thermophila*. Although this family includes other proteins with proposed roles in BB function (i.e., Bbc29p and Bbc39p; Kilburn et al., 2007), it does not include SF-assemblin. Therefore any shared function between DisAp and SF-assemblin may reflect convergent evolution, similar to previous observations for dynamin-like proteins in ciliates and metazoans (Elde et al., 2005). Introduction of TTHERM_00941400 into disA-1 cells rescues BB disorganization (Fig. 1 D). Thus, BB disorganization in *disA-1* is caused by the mutation of DisAp. Moreover, our discovery of DisAp provides proof-of-principle for the combined use of *Tetrahymena* forward genetic screens and next-generation sequencing to identify novel BB mutants.

SF-assemblin is the major component of algal striated fibers (Lechtreck and Melkonian, 1991). Striated fibers in *Paramecium tetraurelia* are composed of multiple polypeptides (Sperling, 1989), suggesting that the family of SF-assemblin genes expanded in ciliates. Indeed, *DISA* belongs to a family of genes that has undergone frequent duplications, especially evident in *P. tetraurelia*, and appears to be conserved in *Giardia*, a lineage distantly related to ciliates (Fig. S1 D and Table S1). However, SF-assemblin from *C. reinhardtii* does not associate with the clade, as assessed by reciprocal BLAST searches that failed to recover identity between these proteins (unpublished data). Therefore, to test whether DisAp localizes to *T. thermophila* KFs, wild-type (WT) GFP-DisAp and mutant disA-1 GFP-disA-1p were localized relative to BBs and KFs in otherwise WT cells. GFP-DisAp localizes to punctae along ciliary rows that colocalize with the proximal portion of the KF (Fig. 1, E and F). In contrast, mutant GFP-disA-1p does not localize to KFs (Fig. 1 E). GFP-DisAp is restricted to ciliary row BBs and is absent from oral apparatus BBs (Fig. 1 E, arrows), which form normally in disA-1 cells (Fig. 1 D, arrows).

DisAp's localization near the base of the KF suggested that it localizes to a discrete domain within the KF. Endogenously tagged DisAp-mCherry also localizes to the proximal portion of the KF (Fig. 1 F). Signal for DisAp is anterior to BBs and decreased below 50% ~500 nm before KF intensity declined to 50%. Consistently, DisAp localized by immuno-EM clustered near the base of the KF (Fig. 1 G, yellow arrows; and Fig. S1 E). We also detected DisAp adjacent to the KF (Fig. 1 G, white arrows), which may reflect a population that has not incorporated into the KF. Thus, DisAp localizes to a domain at the proximal portion of the KF and is phylogenetically distinct from SF-assemblin.

DisAp loss disrupts BB orientation and prevents temperature-induced KF elongation

In disA-1 cells, KFs are disoriented relative to the cellular anterior–posterior axis. Therefore, DisAp could either specify the location of KF attachment to BBs or prevent BB rotation. Normally, KFs from adjacent BBs are aligned along a common axis and are oriented ~180° from postciliary microtubules (Fig. 2 A; Allen, 1969). This KF placement is analogous to vertebrate BBs where the basal foot microtubules are positioned ~180° from the striated rootlet (Steinman, 1968; Peraldi-Roux et al., 1991). In disA-1 cells, KFs are positioned ~180° from the postciliary microtubules (Fig. 2 A), which suggests that DisAp prevents BB rotation and does not affect accessory structure placement. Moreover, disA-1 KFs are shorter than WT KFs (Fig. 2 A). Thus, DisAp prevents BB rotation and is required to establish and/or maintain appropriate KF length. We propose that DisAp functions as a regulator of KF elongation.

BB orientation defects in disA-1 cells are exacerbated by elevated temperature (Fig. 2 B; Jerka-Dziadosz et al., 1995). Thus, short KFs might allow temperature-induced BB rotation. If true, long KFs should prevent BB rotation. One prediction from this inference is that KFs elongate at elevated temperatures to resist BB rotation. To test this, we developed a semiautomated image analysis routine to measure KF length as well as BB orientation, and we assessed these parameters after shifting G1-arrested cells to 37°C and releasing them into the cell

Published December 22, 2014

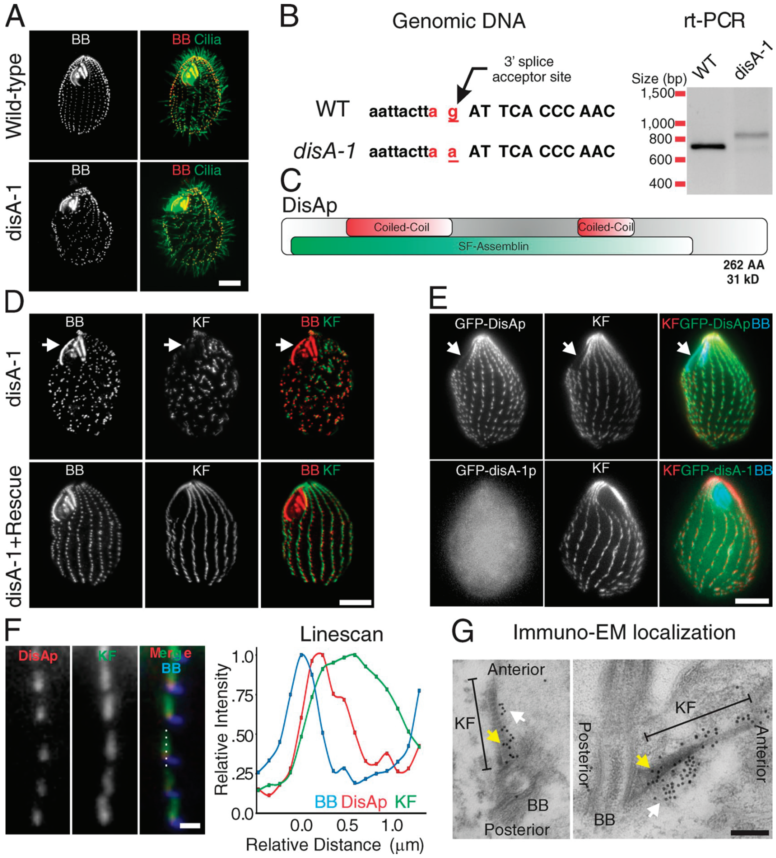


Figure 1. ***DISA* encodes a KF localizing protein.** (A) Disorganized BBs in *disA-1* mutants. BB (centrin; red) and cilia (α-tubulin; green) localization at 30°C is shown. (B) The *disA-1* mutation in Intron 1 of TTHERM_00941400. cDNA size increases due to the retained intron. (C) DisAp domain organization. (D) disA-1 phenotypes at 37°C are rescued with WT *DISA*. Arrows point to the location of the oral apparatus. (E) WT GFP:DisAp and mutant GFP:disA-1p localization relative to KFs and BBs. (F) DisAp-mCherry (red) localizes to the proximal portion of KFs (green). Shown on the right is a fluorescence intensity line scan of a single BB/KF unit. (G) Immuno-EM localization of DisAp-mCherry. Representative transverse (left) and longitudinal (right) sections are taken through a single BB. Yellow arrows point to gold particles associated with KF, and white arrows point to gold particles not associated with the KF. Bars: (A, D, and E) 10 µm; (F) 750 nm; (G) 200 nm.

cycle (Fig. 2 C). Before the temperature shift, WT cells had a mean KF length of 1.10 µm (Fig. 2 D). After the temperature shift, KF elongation reached 1.38 µm in length after 24 h (Fig. 2 D and Fig. S2 A). disA-1 KFs were approximately half as long (0.49 µm), and elongation was more gradual than in WT cells (Fig. 2 D and Fig. S2 B). Unlike WT, disA-1 cells displayed a time-dependent randomization of BB orientation (Fig. 2, E and F; and Fig. S2, C and D). These experiments were performed after a starvation-induced G1 arrest. Because starvation affects cortical organization (Nelsen and Debault, 1978), this could complicate our analyses. However, increased temperature in cycling cells that were not synchronized in G1 also caused WT KF elongation (Fig. 3, A and B) and increased disA-1 BB disorientation (Fig. 3 C). Thus, increased temperature, and not starvation, promotes DisAp-dependent KF elongation and increases the severity of BB disorientation in disA-1 cells. Collectively, these data uncover a novel relationship between KF length and BB orientation. First, the KF is dynamic and elongates in response to elevated temperature. Second, normal KF length requires DisAp. When the KF length is impaired, BBs are susceptible to rotation. We next investigated how temperature induces these changes in BB morphology.

Published December 22, 2014

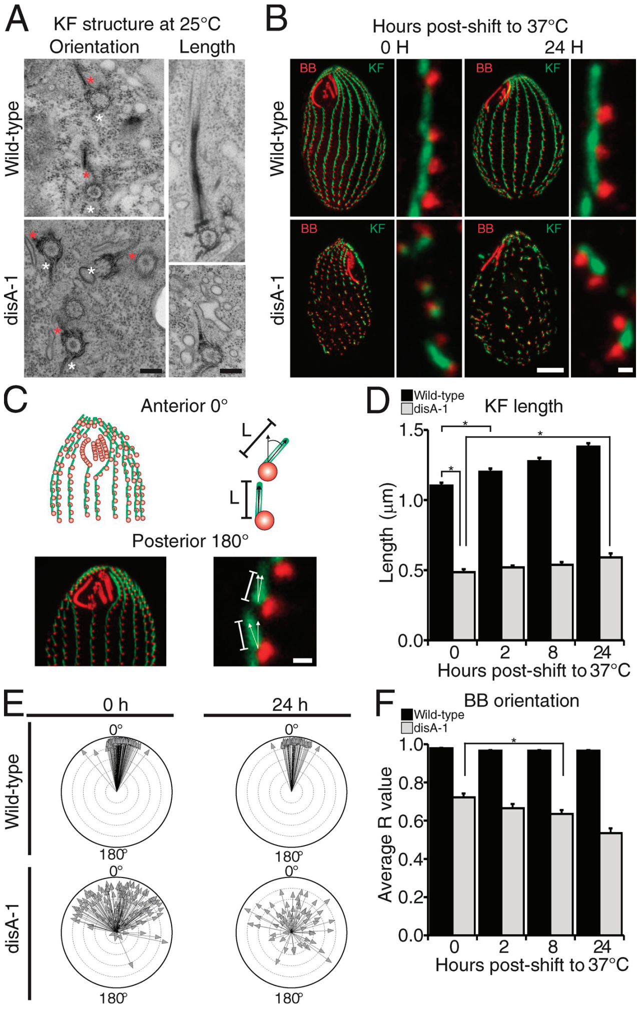


Figure 2. **Elevated temperature lengthens WT KFs and disrupts disA-1 BB orientation.** (A) TEM of ciliary rows at 25°C. DisAp loss causes BBs to rotate and decreases KF length. Red and white asterisks mark KFs and postciliary microtubules, respectively. Bars, 200 nm. (B) BB disorientation increases in disA-1 cells at 37°C (BB, red; KF, green). Bars: (left panels) 10 µm; (enlarged panels) 750 nm. (C) Quantification of KF (green) length and BB (red) orientation. Angular measurements represent the angle between the anterior pole (0°) and the tip of the KF. Length (L) is the distance between the BB and the KF tip. (D) Elevated temperature temporally lengthens KFs. $n > 300$ KFs. (E) BB disorientation in disA-1 cells is temperature sensitive. Arrow direction represents the mean angular measurement (mean vector) for BB orientation within a cell and arrow length represents the R value (mean vector length) for all measured angles for that cell. $n > 100$. (F) Temperature-induced BB disorientation in disA-1 cells. $n > 100$. Brackets indicate the samples being compared and asterisks indicate statistical significance ($P < 0.01$). Error bars indicate SEM.

DisAp confers resistance to mechanical forces produced by ciliary beating

Elevated temperature increases cilia beat frequency and cell swimming speed (Goto et al., 1982; Pearson et al., 2009), which confers greater cilia-generated forces on BBs (Bayless et al., 2012). We explored whether temperature-induced increases in force in disA-1 corresponds with the observed BB disorientation by quantifying cellular swim speeds at differing temperatures (Fig. 3, D and E). At 25°C, WT cells swam at 272 µm/s; this increased to 392 µm/s after a 10-min incubation at 37°C (acute) but decreased to an intermediate level (315 µm/s) after prolonged 24-h incubation at 37°C (chronic). disA-1 cells at 25°C exhibited a reduced swimming rate relative to WT cells (123 µm/s). Acute temperature shift increased the velocity (228 µm/s). However, unlike WT cells, increased motility was not sustained, as chronic maintenance at 37°C decreased the swimming rate below that of disA-1 cells grown at 25°C. This motility defect parallels disA-1 BB disorganization, with prolonged growth at 37°C causing more severe BB disorientation. The initial increase in swim speed in disA-1 cells shifted to 37°C for 10 min (acute) is likely the result of increased beat frequency. However, prolonged exposure to increased beating forces may

Published December 22, 2014

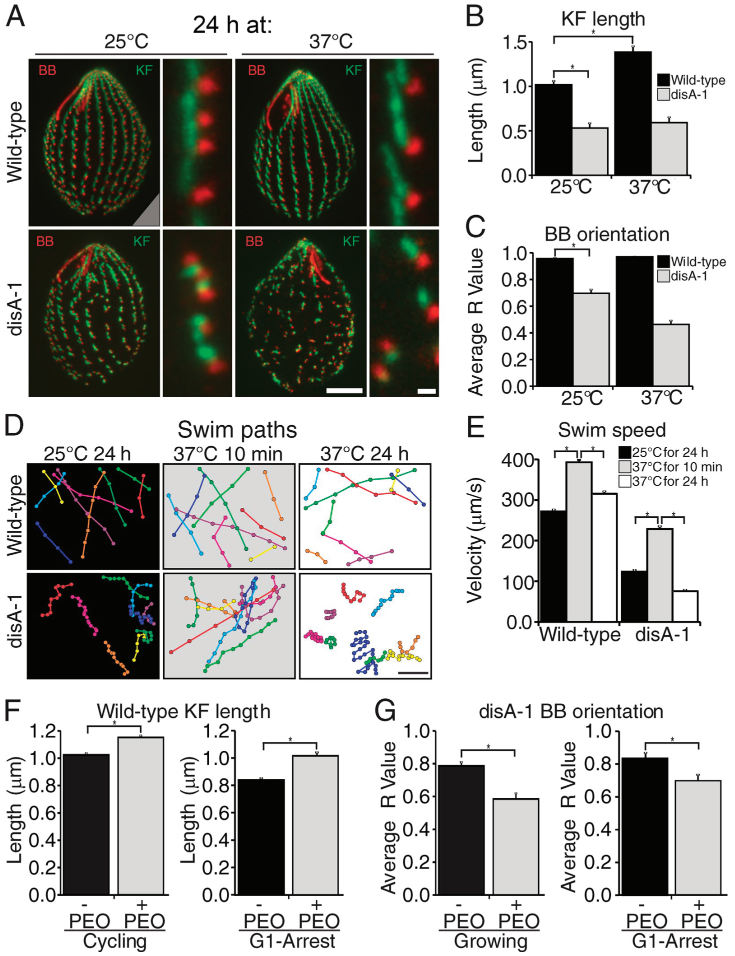


Figure 3. **Increased ciliary forces disorient BBs in disA-1 cells.** (A) Elevated temperature in cycling cells increases the disA-1 phenotype. BB, red; KF, green. Bars: (left panels) 10 µm; (enlarged panels) 750 nm. (B) Elevated temperature lengthens KFs. *n* > 280 KFs. (C) Elevated temperature increases disA-1 BB disorientation. *n* > 54. (D) Elevated temperature increases cell motility. The node spacing represents the distance traveled in 170 ms. Bar, 100 µm. (E) disA-1 cells do not maintain temperature-induced increases in motility. *n* > 48. (F and G) High-viscosity media lengthens WT KFs (*n* > 153 KFs) and disA-1 BB disorientation (*n* > 30) in cycling and G1-arrested cells grown in PEO at 25°C. Brackets indicate the samples being compared, and asterisks indicate statistical significance (P < 0.01). Error bars indicate SEM.

drive BB disorientation, thereby decreasing the effective rate of cell swimming.

Cilia-generated force increases WT KF length and disA-1 BB disorientation

We next tested whether increases in ciliary force influence KF elongation and BB orientation independent of temperature changes. The drag forces (physical resistance) that cilia experience can be increased by increasing their environmental viscosity with polymers (Spoon et al., 1977; Jung et al., 2014). In cycling cells cultured in high viscosity media (polyethylene oxide [PEO]) at 25°C, WT KFs elongated (Fig. 3 F) and disA-1 cells increased BB disorientation (Fig. 3 G). Moreover, high viscosity media also caused G1-arrested WT cells to undergo KF elongation (Fig. 3 F) and disA-1 cells to exhibit randomization of BB orientation (Fig. 3 G). Because G1-arrested cells

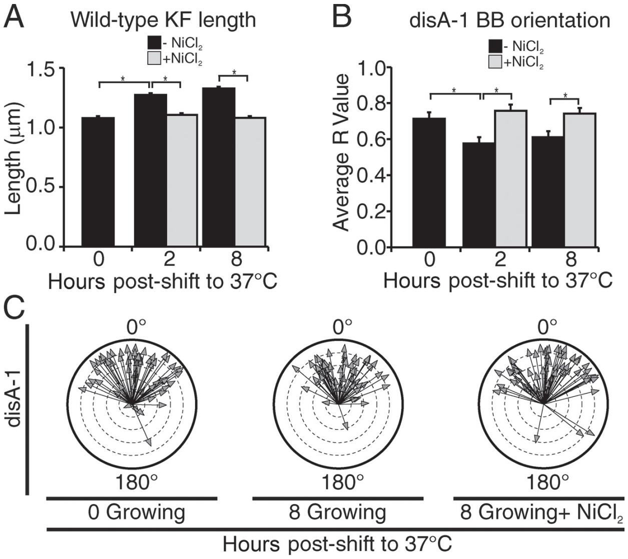


Figure 4. **Decreased cilia beating forces block KF elongation and rescue BB disorientation in disA-1 cells.** (A) Reduced ciliary beating prevents KF elongation. WT cells were grown at 37°C in $NiCl_2$. *n* > 286 KFs. (B) Reduced ciliary beating rescues temperature-induced BB disorientation in disA-1 cells. The circular R value is given for disA-1 cells grown as in A. *n* > 56. (C) Polar plots of the mean angular measurement for disA-1 cells where cilia beating was reduced. *n* > 56. Brackets indicate the samples being compared and asterisks indicate statistical significance (P < 0.01). Error bars indicate SEM.

do not assemble new BBs, BB assembly is not required for KF elongation or BB disorientation. Finally, increasing ciliary beat frequency with the cAMP agonist IBMX (Hennessey and Lampert, 2012) lengthened WT KFs and increased disA-1 BB orientation defects (Fig. S3, B and C), and when high temperature shift was accentuated with increased viscosity, an additive effect was observed (Fig. S3, D–F). Thus, increased ciliary-generated force triggers KF elongation. In the absence of KF elongation, as observed for disA-1, enhanced ciliary forces disrupts BB orientation.

Because cilia-generated force leads to KF elongation, we asked whether a reduction in ciliary beating prevents temperature-induced KF elongation. In the presence of $NiCl_2$ or vanadate, WT temperature-induced KF elongation at 2 and 8 h was abolished (Fig. 4 A and Fig. S3 G), which suggests that KFs elongate due to cilia-generated forces. Consistent with this, growth at 15°C slows cell swimming (Beveridge et al., 2010) and reduces KF length (Fig. S3 G). Moreover, the disorientation observed upon shifting disA-1 cells to 37°C was rescued by reducing ciliary beating with either $NiCl_2$ or vanadate (Fig. 4, B and C). Similarly, growth at 15°C slightly reduced disA-1 BB disorientation (Fig. S3 H; not statistically significant). Thus, cilia-generated force is both necessary and sufficient to increase BB disorientation in disA-1 cells, and BB rotation is resisted by DisAp-mediated KF elongation.

KFs guide nascent BBs and maintain the position of mature BBs

KFs terminate subjacent to the cortical membranes in the membrane-skeletal layer, where stable attachment of the KF may stabilize BBs against rotation. To determine whether KF elongation increases contacts between the KFs and the membrane-skeletal layer, we used three-dimensional image averaging to determine the mean length of the KF that lies near to the cell cortex. This position is defined by centrin, which marks the distal end of the BB near the membrane-skeletal layer (Stemm-Wolf et al., 2005). In WT cells at 25°C, the KF full-width at half maximum (FWHM) intensity above the plane of centrin was 1.09 µm long, which increased to 1.28 µm upon shifting to 37°C for 24 h (Fig. 5 A). In disA-1 cells cultured at 25°C, the KF FWHM was 0.57 µm, and it decreased to 0.51 µm after shifting to 37°C (Fig. 5 B). These results argue that force-dependent KF elongation augments the contact between the KF and anchoring structures in the cell cortex.

BB orientation in ciliates is propagated via a nongenetic process termed cytotaxis, which relies upon preexisting structures, such as old BBs, to constrain the position and orientation of newly arising structures, such as new BBs (Sonneborn, 1964; Beisson and Sonneborn, 1965; Beisson, 2008). Interactions between BBs and KFs or striated rootlets are proposed to organize the even spacing of BBs (Allen, 1969; Wright et al., 1983; Hoops et al., 1984; Lechtreck et al., 2002; Iftode and Fleury-Aubusson, 2003). In ciliates, these interactions occur between the KF and the postciliary microtubules of adjacent BBs in a ciliary row (Fig. 5 C). Nascent BBs are assembled at a mother BB and then transported along the mother BB's KF to separate the daughter from the mother (Fig. 5 C). In disA-1, the association between neighboring BBs and the KF is generally preserved (Fig. 5 D), which suggests that DisAp is not essential to link adjacent BBs to the KF. However, because disA-1 KFs are short and disoriented with respect to the cellular anterior–posterior axis, BB separation along the KF leads to clusters of closely spaced BBs

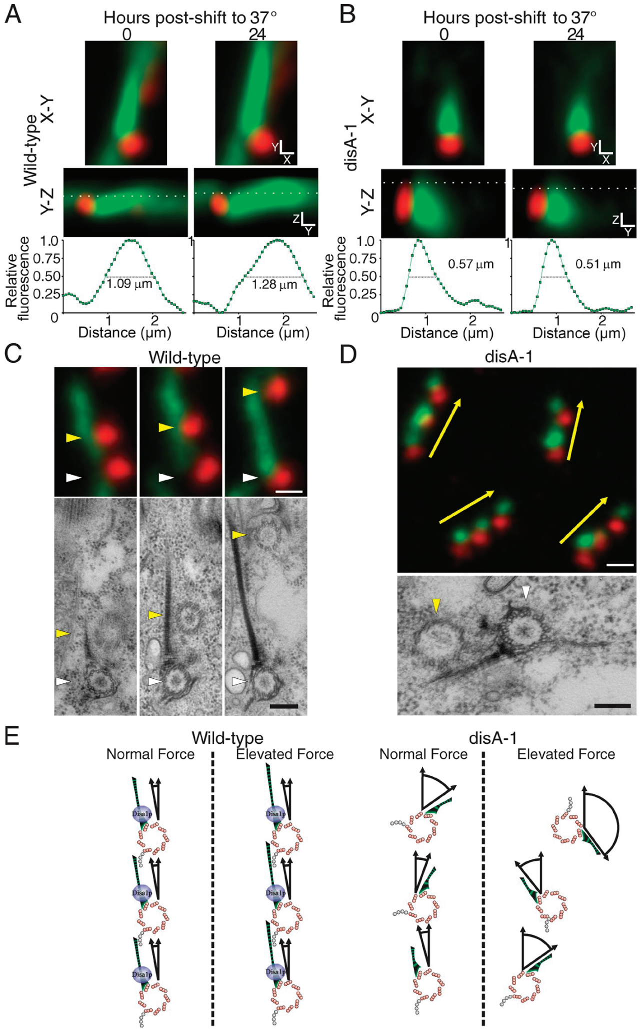


Figure 5. **The KF guides and maintains the position of nascent BBs.** (A and B) Averaged images showing increased KF presence near the membrane-skeletal layer after force-dependent elongation. Shown is the average KF signal (green) at the plane of the cell (broken line) orthogonal to the 90th percentile of centrin BB signal (red). Normalized line scans of fluorescence intensity along the broken line are shown. (C) BBs traverse the KF as they separate from their mother. (C, top) Fluorescence images. White arrowheads, mother BB. Yellow arrowheads, daughter BB. (C, bottom) TEMs representing BB separation. Bar, 200 nm. (D) The association between BBs and KFs is retained in disA-1 cells. (D, top) Shortened and disoriented KFs organize closely spaced BBs aligned along a common, but disoriented, axis (yellow arrows). (D, bottom) EM of a daughter BB (yellow arrowhead) associated with a parent BB (white arrowhead) in a disA-1 cell. (E) WT BBs are oriented along the cellular anterior–posterior axis. When force increases, the KF lengthens in a DisAp-dependent manner, which prevents BB rotation. disA-1 BBs are disorganized, but generally polarize along the cellular anterior–posterior axis. When force increases, KFs do not elongate, leading to force-dependent BB rotation and severe BB disorientation. Bars: (A and B); 320 nm; (C, top) 750 nm; (C, bottom) 200 nm; (D, top) 1.5 µm; (D, bottom) 200 nm.

oriented along a shared axis, each of which deviates from the cellular polarity (Fig. 5 D). Thus, similar to striated fiber-dependent centrosome cohesion and daughter cell positioning (Bahe et al., 2005; Francia et al., 2012), KFs actively position BBs in multiciliary arrays. Moreover, by ensuring that KFs reach an appropriate length, which prevents BB rotation, DisAp allows cytotaxis to perpetuate accurate cortical patterning.

We show that the KF is a major component of the structural environment into which nascent BBs are born. In addition, a genetic input, *DISA*, is required to maintain this environment. We have identified situations (DisAp-deficient) in which cilia-generated forces weaken and partially abolish cytotaxis. This expands upon the concept of structural inheritance to demonstrate that it is both plastic and subservient to the forces that act on BBs.

Conclusion

We demonstrate that the length of the striated fiber in *T. thermophila*, the KF, is responsive to forces generated by cilia. Furthermore, KF elongation stabilizes BB orientation, ensuring

Published December 22, 2014

ciliary alignment and coherent fluid flow. Through next-generation sequencing, we identified *DISA* as a gene responsible for BB organization whose protein is required for KF elongation. Finally, the stability of BB orientation and KF length are important for the propagation of the structural order in cells.

How the forces generated by cilia are sensed and then translated into KF length regulation remains to be determined, and the site of force detection, whether it be the BB or the KF, is also unknown. Because DisAp localizes near the BB and is important for KF elongation, it is an attractive target for force response. Furthermore, our results extend beyond cortical patterning in ciliates. In vertebrates, the striated rootlet, which is analogous to the KF, plays a prominent role in stabilizing the orientation of the ciliary unit (Chien et al., 2013). Therefore, our study raises the intriguing possibility that force sensing and response by BB-associated striated fibers is a conserved mechanism that has independently evolved in different eukaryotic lineages to couple ciliary forces to BB orientation.

Materials and methods

Tetrahymena culture

T. thermophila cells were grown in 2% SPP media (2% proteose peptone, 0.2% glucose, 0.1% yeast extract, and 0.003% Fe-EDTA) at the indicated temperatures (either 15°, 25°C, or 37°C). For all cycling cell studies, cells were analyzed at mid-log phase (density between 10^5 and 4×10^5 cells/ml) as determined using a Coulter Counter Z1 (Beckman Coulter). For starvation experiments, cells were arrested in the G1 phase of the cell cycle by washing and culturing in 10 mM Tris-HCl, pH 7.4, for 18–24 h. For microscopy experiments, analyses were restricted to nondividing cells as judged by those lacking an oral primordium. To expose cells to increased viscosity, *Tetrahymena* were propagated in 2× SPP supplemented with an equal volume of 7.5% polyethylene oxide (mol wt 900,000; Acros Organics), which was prepared in ddH_2O by gentle mixing at 37°C for 24–48 h. Alternatively, 2% SPP was supplemented with 750 nm of the phosphodiesterase inhibitor 3-isobutyl-1-methylxanthine (IBMX; Sigma-Aldrich), which increases cAMP levels and cilia beat frequency. To expose cultures to decreased forces, *Tetrahymena* were propagated in 2% SPP supplemented with 250 µm $NiCl_2$ (Sigma-Aldrich), which was added directly to the culture vessel from a 5-M stock. Dynein-dependent ciliary beating is inhibited by $NiCl_2$, which blocks plasma membrane calcium channels and directly inhibits dynein motors (Larsen and Satir, 1991). Alternatively, 2% SPP was supplemented with 750 µm sodium orthovanadate (Sigma-Aldrich; Gibbons et al., 1978; Nilsson, 1999), which was added directly to the culture vessel from a 100-mm stock (pH 10).

Plasmids and *Tetrahymena* strain construction

To rescue the disA-1 phenotype with WT *DISA*, disA-1 cells (IA217) were transformed with the *DISA* ORF, and flanking sequences were inserted exogenously into the genome at *RPL29*. Specifically, *DISA* was PCR amplified from WT (B2086) cells using oligos (5′-CGCTGCAGAAAGATAGATGCTT-GCTTGC-3′ and 5′-CGGAGCTCGCTGTATTCTAAAGTTCAAG-3′) and cloned into pBSMCSCHX. After release with Blp1 digest, the rescue construct was biolistically transformed into disA-1 cells and selected for cycloheximide resistance and rescue of the disA-1 phenotype.

GFP-DISA and GFP-disA-1 were cloned into pNEO2-MTTpr-GFP that inserts an MTT-GFP cassette upstream of the endogenous gene (Winey et al., 2012). The cassette containing either WT or *disA-1* mutant sequence was inserted into otherwise WT cells. 0.6 kb of sequence upstream of *DISA* was PCR amplified (5′-CGGAGCTCCTGTTTAAAATTAAGCATGCTC-3′ and 5′-CGGGATTCGCCTTATTAACCGTTTCCTT-3′) and cloned into pNEO2-MTTpr-GFP. Next, a 0.6-kb sequence of either DISA or disA-1 was PCR amplified (5′-CGCTCGAGATGTCTGCTTTCGGCTCTCC-3′ and 5′-CGGGGCCCCTTAATTTCTTTACCCTTTC-3′) and cloned into the plasmid to produce either pNEO2-MTT-GFP-DISA or pNEO2-MTT-GFP-disA-1.

A DisAp-mCherry strain was constructed. The p4T2-1:DisA:mCherry cassette integrates at the endogenous *DISA* locus and remains under the control of the endogenous promoter. p4T2-1:DISA:mCherry was generated by PCR amplifying (5′-CGGGATCCAAAATTGCAACTATCTAAAC-3′ and 5′-CGGAGCTCTCAAGTGAGCTTTAACTATC-3′) and cloning the final 0.6 kb of *DISA* without the TGA stop codon into p4T2-1-mCherryLAP (Winey et al., 2012). A 0.9-kb fragment downstream of the TGA stop codon (5′-CGGGTACCAGAAAATCATATTGAAACAC-3′ and 5′-CGGAATTCTT-GACGGAAGTTCTTATCAATTCTCTAGCAAGTG-3′) was then cloned into the plasmid to create p4T2-1:DISA:mCherry. This plasmid contains NEO2 drug selection.

Next-generation sequencing and identification of disA-1

To generate a backcross of the *disA-1* mutation, IA217 (*disA-1*) and B1868 (WT) *Tetrahymena* lines were crossed to produce micronuclear heterozygous F1 progeny. Two F1 clones of different mating types (F1.1 and F1.8) were then mated to produce 18 F2 *disA-1* mutant lines. Total genomic DNA was purified from each line using a urea-SDS lysis and phenol:chloroform extraction (Gaertig et al., 1994) and measured using a Qubit Fluorometer (Invitrogen). Equal DNA was pooled for all 18 F2 clones to produce a single Illumina Tru-seq library. A second DNA preparation of B1868 was also prepared. Using Illumina's standard TruSeq DNA library preparation, DNA was first sheared to a 300–400-bp distribution using sonication, and then end-repaired and A-tailed using a combination of T4 polynucleotide kinase (PNK), T4 DNA polymerase, and Klenow. A-tailed DNA was then ligated to Illumina Truseq adapters, and further independently indexed using PCR. Both libraries were size selected to remove adapter and PCR dimers, pooled, and co-sequenced on a single lane of the Illumina HiSeq 2000 using a 2 × 100 base pair format. Postsequencing bioinformatics were used to separate sequences from each of the two libraries using the unique indexes incorporated during the library process.

The pipeline for mapping and variant calling in the data are reported in a publicly available Wiki (https://github.com/jewmanchue/pooled-mapping). To improve read quality, 10 and 5 bp were trimmed from the 5′ and 3′ ends, respectively. Reads were locally aligned simultaneously to two reference sequences (micronucleus: tetrahymena_thermophila_sb210__mic__2_supercontigs; and macronucleus: JCVI-TTA1-2.2) using bowtie-2.0.0. (Langmead and Salzberg, 2012). Discordant mate pairs were discarded. The reads mapped to the genome were >98% for B1868 control and >95% for the *disA-1* F2. Sequencing coverage for the entire genome was 0.96 and 0.60 for the B1868 and *disA-1* strains, respectively. Alignment cleanup was performed using Samtools to remove PCR duplicates GenomeAnalysisTK-2.4-7 was used to realign indels and polish alignments (Li et al., 2009; McKenna et al., 2010; DePristo et al., 2011). Joint variant calling for the *disA-1* F2 and the B1868 control pool was performed using SNVerPool (Wei et al., 2011).

Heuristic filters were applied to the 8,795,723 mutations (SNVs, 7,346,955; insertions, 648,913; deletions, 799,855) using the Genotype-Phenotype Association Toolkit (https://github.com/jewmanchue/vcflib/wiki). Positions in the genome with a depth <5 in the B1868 or *disA-1* mutant pools were removed. The second filter removed positions where the B1868 pool contained the nonreference allele. The last filter leniently enforced a recessive model by requiring the frequency of the nonreference allele in the *disA-1* mutant pool to be >0.75. A filter of 0.75 was used to avoid false negatives created from sequencing errors. However, the *disA-1* mutation did have an allele frequency of 1.0. After filtering, only 206 mutations were remaining, 26 of which were homozygous nonreference (allele frequency of one) in the *disA-1* mutant pool (Fig. S1 C). We focused on the nine mutations mapped to the macronuclear sequence. Nine candidate positions for the *disA-1* mutation were identified and narrowed down by searching for proteins containing domains commonly associated with BBs and their auxiliary structures. Hand annotation revealed that one of the mutations was a splice site mutation in the 3′ acceptor site (G to A) in TTHERM_00941400. This mutation was found in both the macronuclear and micronuclear sequences. The mutation was also confirmed by both PCR of the genomic region and cDNA of TTHERM_00941400 and sequencing.

Phylogenetic analysis

The amino acid sequence of DisAp from *T. thermophila* (XP_001026900) was used in a protein–protein blast query (http://www.ncbi.nlm.nih.gov) to identify related sequences (Table S1). Alignments were generated with ClustalW2 (http://www.ebi.ac.uk) and trimmed by eye to eliminate insertions and deletions. Model fitting and tree inference of the alignment was performed with Mr. Bayes v3.2.2 (Ronquist and Huelsenbeck, 2003). The "rtrev" amino acid substitution model was best supported by the data and used for inferring the tree. After the burn-in phase, the remainder of 500,000 generations of Markov chain Monte Carlo analysis were considered for

Published December 22, 2014

inference of the tree. The 50% majority rule consensus tree was generated and visualized with FigTree v1.4.

Immunocytochemistry

For immuno-cytochemical analyses, 1–3 × 10^5 cells were pelleted at 1,500 *g* in a 1.5-ml Eppendorf tube and fixed for 20–30 min with 1.5 ml of 70% ethanol + 0.2% Triton X-100. Cells were washed with 10 mM Tris-buffered saline and blocked overnight at 4°C in 1% BSA in 10 mM TBS. Cells were immunostained by incubating overnight at 4°C in primary antibody (mouse anti-KF [5D8], 1:400; Jerka-Dziadosz et al., 1995; rabbit anti-centrin, 1:2,000, a gift from A. Stemm-Wolf and M. Winey, University of Colorado Boulder, Boulder, CO; Stemm-Wolf et al., 2005; rabbit anti–α tubulin, DM1a; Sigma-Aldrich) followed by a 1-h incubation at room temperature in secondary antibody (goat anti–mouse Alexa Fluor 594, 1:2,000; goat anti–rabbit Alexa Fluor 488, 1:2,000; goat anti–Alexa Fluor 647, 1:2,000; Invitrogen). Cells were mounted in Citifluor mounting media (Citifluor LTD) using #1.5 coverslips and sealed with nail polish. All antibodies were diluted in 1% BSA/TBS. Cells were washed (3 × 5 min) with 1% BSA/TBS after primary and secondary antibody incubations.

Light microscopy

For the localization experiments in Fig. 1, an inverted microscope (Ti Eclipse; Nikon) with a 100× Plan-Apochromat (NA 1.4) objective lens (Nikon) was used. Images were captured with an electron-multiplying charge-coupled device (EMCCD) 888E camera (iXon; Andor Technology). For all other experiments, confocal microscopy was performed using an inverted microscope (Ti Eclipse) with a 100× Plan-Apochromat (NA 1.43) objective lens (Nikon) and a Swept Field confocal scan head (Prairie Technologies). Confocal images were acquired in slit mode with a slit size of 35 µm and a z-step size of 200 nm, and detected with a charge-coupled device (CCD) camera (Clara; Andor Technology). Images were acquired with Elements software (Nikon) and all fixed cells were imaged at room temperature.

Transmission EM (TEM)

EM was performed as described previously (Pearson et al., 2009; Bayless et al., 2012). A *Tetrahymena* strain expressing endogenous C-terminal DisAp-mCherry was grown to mid-log phase and then prepared for immuno-EM using high-pressure freezing and freeze substitution (HPF-FS; Dahl and Staehelin, 1989; Meehl et al., 2009). *T. thermophila* cells were pelleted, high-pressure frozen (HPM-010; Bal-Tec), freeze substituted in 0.25% glutaraldehyde/0.1% uranyl acetate in acetone, and embedded in Lowicryl HM20. 60-nm serial sections were cut and put on nickel slot grids, blocked with 1% milk in PBS–Tween 20, and incubated with anti-mCherry (rabbit polyclonal; a gift from I. Cheeseman, Massachusetts Institute of Technology, Cambridge, MA) at 1:100. 15 nm of gold-conjugated secondary antibody was applied to the grids at a dilution of 1:20 (Ted Pella). Grids were poststained with 2% uranyl acetate and lead citrate.TtBld10 was then localized in 60-nm sections using TEM. Images were collected using a Philips CM10 electron microscope (Philips) equipped with a Gatan BioScan2 CCD camera (Gatan). For structural analyses of *disA-1* BB defects, *disA-1* and control (B1868) cells were subjected to HPF-FS after growth at 25°C.

***Tetrahymena* motility measurements**

Free-swimming *Tetrahymena* in glass-bottom dishes were imaged using a 20× objective lens (pixel size, 330 nm) and transmitted light on the confocal microscope (see "Light microscopy"). For each field of view, images were captured at ~170-ms intervals for a total of 30 s. To track motility paths, we marked the anterior tip of individual *Tetrahymena* that displayed directional motility for the duration of their swim path while they remained in focus. Care was taken to avoid *Tetrahymena* at the glass surface. Tracking analysis was facilitated by the MTrackJ (Meijering et al., 2012) plugin bundled with FIJI (Schindelin et al., 2012).

Image analysis: KF length and BB orientation

KF length and BB orientation quantification were performed in a semiautomated fashion using the macro scripting language and plugins contained within the FIJI build of ImageJ. Image stacks were preprocessed with a Laplacian of Gaussian filter (LOG; radius, 1 pixel) to reduce noise and enhance feature edges. 32-bit LOG stacks were inverted, their contrast was adjusted (minimum = mode pixel intensity + (1/2) × standard deviation pixel intensity; maximum = maximum pixel intensity), and the images were merged to create an 8-bit RGB image stack. To quantify KF length for individual cells, 10 KFs with clear separation from neighboring KFs were manually measured by tracing with the freehand line tool. To quantify BB orientation, a box (10 µm wide × 5 µm tall) was placed in the center of a *Tetrahymena* cell and angular measurements were made for 10 BBs within that box. For each BB, the angular measurement represents the angle between the tip of the KF and the anterior pole of the cell (Fig. 2 C). For each cell, the mean vector and the length of the mean vector (R value) for the 10 measured BBs were calculated using circular statistics and displayed on polar plots. Each cell was measured twice, once on each side; thus each cell produced two R values. On the polar plots the mean vector is represented by the direction of the arrow. Variance in the mean vector is represented by the length of the arrow. An arrow length of 1 indicates no variation in the mean vector and an arrow length of 0 represents pure randomness in the mean vector. On the polar plots, the dashed circles represent 0.2 arbitrary units of R value. To compare the amount of BB orientation defects across different populations of cells, the mean R value for the cell population was determined in linear space. Circular statistics were calculated using the ORIANA circular statistics suite (Kovach Computing Services).

Image analysis: fluorescent image averaging

The brightest centrin (BB) voxel for an individual BB was determined. A 5-µm box was centered over this voxel in the x, y, and z dimensions. The raw BB and KF image stacks were cropped in the xy dimension using the 5-µm box, and they were cropped in the z dimension by taking five slices below the slice containing the brightest BB voxel and five slices above the brightest BB voxel (11 slices total; 3.3 µm). Next, cropped stacks were rotated so that the tip of the KF was aligned with a straight line that ran down the middle of the 5-µm box and passed through the brightest centrin pixel. This procedure was performed on 100 BBs from 10 different cells for each condition. The raw images used for averaging were part of the 0 h and 24 h time points (SPP condition) of the dataset used in Fig. S3 (D and E). To create the average image stack, individual image stacks were averaged on a per-slice basis. All image cropping was performed with FIJI using the crop, rotate, and duplicate stack commands. The yz images were created by rotating the averaged image stacks in three dimensions using the TransformJ plugin.

Statistical analysis

All linear statistical analyses were performed in Excel (Microsoft). All tests for significance were unpaired, two-tailed *t* tests. All error bars indicate SEM. Statistical significance was set at P < 0.01.

Online supplemental material

Fig. S1 shows the scheme used for the identification of the *disA-1* mutation, a phylogenetic tree of related DisAp proteins, and immuno-EM confirming DisAp's localization at the KF. Fig. S2 shows the frequency distributions of WT and disA-1 KF length and BB orientation upon temperature shift, which documents population-wide shifts in KF length and BB orientation. Fig. S3 shows that WT BBs are resistant to force-induced orientation defects, whereas additional force perturbations impact WT KF length and disA-1 BB orientation. Table S1 lists the *Tetrahymena DISA-1* clade members. Online supplemental material is available at http://www.jcb.org/cgi/content/full/jcb.201409123/DC1.

We thank Alex Stemm-Wolf and Dr. Brian Mitchell for comments on the manuscript. We thank Dr. Ken Jones, Joe Brown, Dr. Bifeng Gao, and Katrina Diener for providing sequencing and preliminary analyses.

The sequencing was supported by a pilot grant from the University of Colorado Denver Genomics and Microarray Core made possible by the Genomics and Biostatistics/Bioinformatics Shared Resources of Colorado's National Institutes of Health (NIH)/National Cancer Institute Cancer Center Support Grant P30CA046934. C.G. Pearson is supported by NIH-NIGMS GM0099820 and the Boettcher Webb-Waring Foundation. C.G. Pearson and N.C. Elde are supported by the Pew Biomedical Scholars Program.

The authors declare no competing financial interests.

Submitted: 25 September 2014
Accepted: 25 November 2014

References

Allen, R.D. 1967. Fine structure, reconstruction and possible functions of components of the cortex of *Tetrahymena pyriformis*. *J. Protozool.* 14:553–565. http://dx.doi.org/10.1111/j.1550-7408.1967.tb02042.x

Allen, R.D. 1969. The morphogenesis of basal bodies and accessory structures of the cortex of the ciliated protozoan *Tetrahymena pyriformis*. *J. Cell Biol.* 40:716–733. http://dx.doi.org/10.1083/jcb.40.3.716

Bahe, S., Y.-D. Stierhof, C.J. Wilkinson, F. Leiss, and E.A. Nigg. 2005. Rootletin forms centriole-associated filaments and functions in centrosome cohesion. *J. Cell Biol.* 171:27–33. http://dx.doi.org/10.1083/jcb.200504107

Bayless, B.A., T.H. Giddings Jr., M. Winey, and C.G. Pearson. 2012. Bld10/Cep135 stabilizes basal bodies to resist cilia-generated forces. *Mol. Biol. Cell*. 23:4820–4832. http://dx.doi.org/10.1091/mbc.E12-08-0577

Beisson, J. 2008. Preformed cell structure and cell heredity. *Prion*. 2:1–8. http://dx.doi.org/10.4161/pri.2.1.5063

Beisson, J., and T.M. Sonneborn. 1965. Cytoplasmic inheritance of the organization of the cell cortex in *Paramecium aurelia. Proc. Natl. Acad. Sci. USA*. 53:275–282. http://dx.doi.org/10.1073/pnas.53.2.275

Beveridge, O.S., O.L. Petchey, and S. Humphries. 2010. Mechanisms of temperature-dependent swimming: the importance of physics, physiology and body size in determining protist swimming speed. *J. Exp. Biol.* 213:4223–4231. http://dx.doi.org/10.1242/jeb.045435

Chien, Y.-H., M.E. Werner, J. Stubbs, M.S. Joens, J. Li, S. Chien, J.A.J. Fitzpatrick, B.J. Mitchell, and C. Kintner. 2013. Bbof1 is required to maintain cilia orientation. *Development*. 140:3468–3477. http://dx.doi.org/10.1242/dev.096727

Dahl, R., and L.A. Staehelin. 1989. High-pressure freezing for the preservation of biological structure: theory and practice. *J. Electron Microsc. Tech.* 13:165–174. http://dx.doi.org/10.1002/jemt.1060130305

DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491–498. http://dx.doi.org/10.1038/ng.806

Elde, N.C., G. Morgan, M. Winey, L. Sperling, and A.P. Turkewitz. 2005. Elucidation of clathrin-mediated endocytosis in *tetrahymena* reveals an evolutionarily convergent recruitment of dynamin. *PLoS Genet.* 1:e52. http://dx.doi.org/10.1371/journal.pgen.0010052

Francia, M.E., C.N. Jordan, J.D. Patel, L. Sheiner, J.L. Demerly, J.D. Fellows, J.C. de Leon, N.S. Morrissette, J.-F. Dubremetz, and B. Striepen. 2012. Cell division in Apicomplexan parasites is organized by a homolog of the striated rootlet fiber of algal flagella. *PLoS Biol.* 10:e1001444. http://dx.doi.org/10.1371/journal.pbio.1001444

Frankel, J. 1979. An analysis of cell-surface patterning in *Tetrahymena. Symp. Soc. Dev. Biol.* 37:215–246.

Frankel, J. 2000. Cell biology of *Tetrahymena thermophila. Methods Cell Biol.* 62:27–125. http://dx.doi.org/10.1016/S0091-679X(08)61528-9

Frankel, J. 2008. What do genic mutations tell us about the structural patterning of a complex single-celled organism? *Eukaryot. Cell*. 7:1617–1639. http://dx.doi.org/10.1128/EC.00161-08

Gaertig, J., T.H. Thatcher, L. Gu, and M.A. Gorovsky. 1994. Electroporation-mediated replacement of a positively and negatively selectable β-tubulin gene in *Tetrahymena thermophila. Proc. Natl. Acad. Sci. USA*. 91:4549–4553. http://dx.doi.org/10.1073/pnas.91.10.4549

Gibbons, I.R., M.P. Cosson, J.A. Evans, B.H. Gibbons, B. Houck, K.H. Martinson, W.S. Sale, and W.J. Tang. 1978. Potent inhibition of dynein adenosinetriphosphatase and of the motility of cilia and sperm flagella by vanadate. *Proc. Natl. Acad. Sci. USA*. 75:2220–2224. http://dx.doi.org/10.1073/pnas.75.5.2220

Gibbons, I.R. 1981. Cilia and flagella of eukaryotes. *J. Cell Biol.* 91:107s–124s. http://dx.doi.org/10.1083/jcb.91.3.107s

Goto, M., K. Ohki, and Y. Nozawa. 1982. Evidence for a correlation between swimming velocity and membrane fluidity of *Tetrahymena* cells. *Biochim. Biophys. Acta*. 693:335–340. http://dx.doi.org/10.1016/0005-2736(82)90440-0

Hennessey, T.M., and T.J. Lampert. 2012. Behavioral bioassays and their uses in *Tetrahymena. Methods Cell Biol.* 109:393–410. http://dx.doi.org/10.1016/B978-0-12-385967-9.00015-3

Hoops, H.J., R.L. Wright, J.W. Jarvik, and G.B. Witman. 1984. Flagellar waveform and rotational orientation in a *Chlamydomonas* mutant lacking normal striated fibers. *J. Cell Biol.* 98:818–824. http://dx.doi.org/10.1083/jcb.98.3.818

Iftode, F., and A. Fleury-Aubusson. 2003. Structural inheritance in *Paramecium*: ultrastructural evidence for basal body and associated rootlets polarity transmission through binary fission. *Biol. Cell*. 95:39–51. http://dx.doi.org/10.1016/S0248-4900(03)00005-4

Jana, S.C., G. Marteil, and M. Bettencourt-Dias. 2014. Mapping molecules to structure: unveiling secrets of centriole and cilia assembly with near-atomic resolution. *Curr. Opin. Cell Biol.* 26:96–106. http://dx.doi.org/10.1016/j.ceb.2013.12.001

Jerka-Dziadosz, M., L.M. Jenkins, E.M. Nelsen, N.E. Williams, R. Jaeckel-Williams, and J. Frankel. 1995. Cellular polarity in ciliates: persistence of global polarity in a disorganized mutant of *Tetrahymena thermophila* that disrupts cytoskeletal organization. *Dev. Biol.* 169:644–661. http://dx.doi.org/10.1006/dbio.1995.1176

Jung, I., T.R. Powers, and J.M. Valles Jr. 2014. Evidence for two extremes of ciliary motor response in a single swimming microorganism. *Biophys. J.* 106:106–113. http://dx.doi.org/10.1016/j.bpj.2013.11.3703

Kilburn, C.L., C.G. Pearson, E.P. Romijn, J.B. Meehl, T.H. Giddings Jr., B.P. Culver, J.R. Yates III, and M. Winey. 2007. New *Tetrahymena* basal body protein components identify basal body domain structure. *J. Cell Biol.* 178:905–912. http://dx.doi.org/10.1083/jcb.200703109

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 9:357–359. http://dx.doi.org/10.1038/nmeth.1923

Larsen, J., and P. Satir. 1991. Analysis of Ni$^{(2+)}$-induced arrest of *Paramecium* axonemes. *J. Cell Sci.* 99:33–40.

Lechtreck, K.F., and M. Melkonian. 1991. Striated microtubule-associated fibers: identification of assemblin, a novel 34-kD protein that forms paracrystals of 2-nm filaments in vitro. *J. Cell Biol.* 115:705–716. http://dx.doi.org/10.1083/jcb.115.3.705

Lechtreck, K.-F., and M. Melkonian. 1998. SF-assemblin, striated fibers, and segmented coiled coil proteins. *Cell Motil. Cytoskeleton*. 41:289–296. http://dx.doi.org/10.1002/(SICI)1097-0169(1998)41:4<289::AID-CM2>3.0.CO;2-1

Lechtreck, K.-F., J. Rostmann, and A. Grunow. 2002. Analysis of *Chlamydomonas* SF-assemblin by GFP tagging and expression of antisense constructs. *J. Cell Sci.* 115:1511–1522.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352

Marshall, W.F., and C. Kintner. 2008. Cilia orientation and the fluid mechanics of development. *Curr. Opin. Cell Biol.* 20:48–52. http://dx.doi.org/10.1016/j.ceb.2007.11.009

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303. http://dx.doi.org/10.1101/gr.107524.110

Meehl, J.B., T.H. Giddings Jr., and M. Winey. 2009. High pressure freezing, electron microscopy, and immuno-electron microscopy of *Tetrahymena thermophila* basal bodies. *Methods Mol. Biol.* 586:227–241. http://dx.doi.org/10.1007/978-1-60761-376-3_12

Meijering, E., O. Dzyubachyk, and I. Smal. 2012. Methods for cell and particle tracking. *Methods Enzymol.* 504:183–200. http://dx.doi.org/10.1016/B978-0-12-391857-4.00009-4

Nelsen, E.M., and L.E. Debault. 1978. Transformation in *Tetrahymena pyriformis*: description of an inducible phenotype. *J. Protozool.* 25:113–119. http://dx.doi.org/10.1111/j.1550-7408.1978.tb03880.x

Nilsson, J.R. 1999. Vanadate affects nuclear division and induces aberrantly-shaped cells during subsequent cytokinesis in *Tetrahymena. J. Eukaryot. Microbiol.* 46:24–33. http://dx.doi.org/10.1111/j.1550-7408.1999.tb04580.x

Park, T.J., B.J. Mitchell, P.B. Abitua, C. Kintner, and J.B. Wallingford. 2008. Dishevelled controls apical docking and planar polarization of basal bodies in ciliated epithelial cells. *Nat. Genet.* 40:871–879. http://dx.doi.org/10.1038/ng.104

Pearson, C.G. 2014. Choosing sides—asymmetric centriole and basal body assembly. *J. Cell Sci.* 127:2803–2810. http://dx.doi.org/10.1242/jcs.151761

Pearson, C.G., D.P.S. Osborn, T.H. Giddings Jr., P.L. Beales, and M. Winey. 2009. Basal body stability and ciliogenesis requires the conserved component Poc1. *J. Cell Biol.* 187:905–920. http://dx.doi.org/10.1083/jcb.200908019

Peraldi-Roux, S., C. Klotz, B. Nguyen-Thanh-Dao, and J. Gabrion. 1991. A common epitope is shared by ciliary rootlets and cell-cell adherens junctions in ciliated ependymal cells. *J. Cell Sci.* 99:297–306.

Rayner, C.F., A. Rutman, A. Dewar, M.A. Greenstone, P.J. Cole, and R. Wilson. 1996. Ciliary disorientation alone as a cause of primary ciliary dyskinesia syndrome. *Am. J. Respir. Crit. Care Med.* 153:1123–1129. http://dx.doi.org/10.1164/ajrccm.153.3.8630555

Ronquist, F., and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19:1572–1574. http://dx.doi.org/10.1093/bioinformatics/btg180

Salisbury, J.L., A. Baron, B. Surek, and M. Melkonian. 1984. Striated flagellar roots: isolation and partial characterization of a calcium-modulated contractile organelle. *J. Cell Biol.* 99:962–970. http://dx.doi.org/10.1083/jcb.99.3.962

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al. 2012. Fiji: an open-source platform for biological-image analysis. *Nat. Methods*. 9:676–682. http://dx.doi.org/10.1038/nmeth.2019

Sonneborn, T.M. 1964. The differentiation of cells. *Proc. Natl. Acad. Sci. USA*. 51:915–929. http://dx.doi.org/10.1073/pnas.51.5.915

Sperling, L. 1989. Isolation and partial characterization of ciliary rootlets from *Paramecium tetraurelia*. *In* Cytoskeletal and Extracellular Proteins. P.D.U. Aebi, and P.D.J. Engel, editors. Springer Berlin, Heidelberg. 316–318.

Sperling, L., G. Keryer, F. Ruiz, and J. Beisson. 1991. Cortical morphogenesis in *Paramecium*: a transcellular wave of protein phosphorylation involved in ciliary rootlet disassembly. *Dev. Biol.* 148:205–218. http://dx.doi.org/10.1016/0012-1606(91)90330-6

Spoon, D.M., I.I. Feise CO, and R.S. Youn. 1977. Poly(ethylene oxide), a new slowing agent for protozoa. *J. Protozool.* 24:471–474. http://dx.doi.org/10.1111/j.1550-7408.1977.tb04779.x

Steinman, R.M. 1968. An electron microscopic study of ciliogenesis in developing epidermis and trachea in the embryo of *Xenopus laevis*. *Am. J. Anat.* 122:19–55. http://dx.doi.org/10.1002/aja.1001220103

Stemm-Wolf, A.J., G. Morgan, T.H. Giddings Jr., E.A. White, R. Marchione, H.B. McDonald, and M. Winey. 2005. Basal body duplication and maintenance require one member of the *Tetrahymena thermophila* centrin gene family. *Mol. Biol. Cell.* 16:3606–3619. http://dx.doi.org/10.1091/mbc.E04-10-0919

Tamm, S.L., T.M. Sonneborn, and R.V. Dippell. 1975. The role of cortical orientation in the control of the direction of ciliary beat in *Paramecium*. *J. Cell Biol.* 64:98–112. http://dx.doi.org/10.1083/jcb.64.1.98

Wei, Z., W. Wang, P. Hu, G.J. Lyon, and H. Hakonarson. 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 39:e132. http://dx.doi.org/10.1093/nar/gkr599

Winey, M., A.J. Stemm-Wolf, T.H. Giddings Jr., and C.G. Pearson. 2012. Cytological analysis of *Tetrahymena thermophila*. *Methods Cell Biol.* 109:357–378. http://dx.doi.org/10.1016/B978-0-12-385967-9.00013-X

Wloga, D., and J. Frankel. 2012. From molecules to morphology: cellular organization of *Tetrahymena thermophila*. *Methods Cell Biol.* 109:83–140. http://dx.doi.org/10.1016/B978-0-12-385967-9.00005-0

Wright, R.L., B. Chojnacki, and J.W. Jarvik. 1983. Abnormal basal-body number, location, and orientation in a striated fiber-defective mutant of *Chlamydomonas reinhardtii*. *J. Cell Biol.* 96:1697–1707. http://dx.doi.org/10.1083/jcb.96.6.1697

Yang, J., X. Liu, G. Yue, M. Adamian, O. Bulgakov, and T. Li. 2002. Rootletin, a novel coiled-coil protein, is a structural component of the ciliary rootlet. *J. Cell Biol.* 159:431–440. http://dx.doi.org/10.1083/jcb.200207153

Galati et al., http://www.jcb.org/cgi/content/full/jcb.201409123/DC1

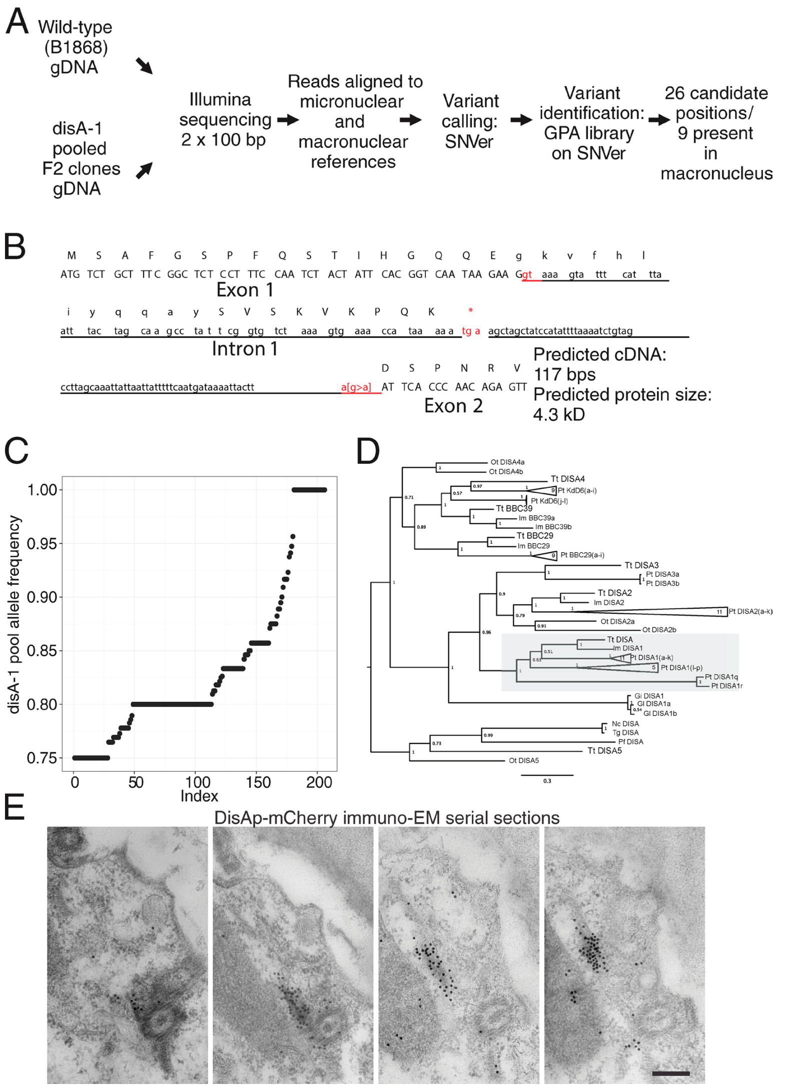

A
Wild-type (B1868) gDNA
disA-1 pooled F2 clones gDNA
Illumina sequencing 2 x 100 bp
Reads aligned to micronuclear and macronuclear references
Variant calling: SNVer
Variant identification: GPA library on SNVer
26 candidate positions/ 9 present in macronucleus
B
M S A F G S P F Q S T I H G Q Q E g k v f h l
ATG TCT GCT TTC GGC TCT CCT TTC CAA TCT ACT ATT CAC GGT CAA TAA GAA Ggt aaa gta ttt cat tta
Exon 1
i y q q a y S V S K V K P Q K *
att tac tag caa gcc tat tcg gtg tct aaa gtg aaa cca taa aaa tga agctagctatccatattttaaaatctgtag
Intron 1
ccttagcaaattattaattatttttcaatgataaaattactt a[g>a] AT TCA CCC AAC AGA GTT
D S P N R V
Exon 2
Predicted cDNA: 117 bps
Predicted protein size: 4.3 kD
C
disA-1 pool allele frequency
1.00
0.95
0.90
0.85
0.80
0.75
0
50
100
150
200
Index
D
Ot DISA4a
Ot DISA4b
Tt DISA4
9 Pt KdD6(a-i)
Pt KdD6(j-l)
Tt BBC39
Im BBC39a
Im BBC39b
Tt BBC29
Im BBC29
9 Pt BBC29(a-i)
Tt DISA3
Pt DISA3a
Pt DISA3b
Tt DISA2
Im DISA2
11 Pt DISA2(a-k)
Ot DISA2a
Ot DISA2b
Tt DISA
Im DISA1
11 Pt DISA1(a-k)
5 Pt DISA1(l-p)
Pt DISA1q
Pt DISA1r
Gi DISA1
Gl DISA1a
Gl DISA1b
Nc DISA
Tg DISA
Pf DISA
Tt DISA5
Ot DISA5
0.71
0.97
0.57
0.89
0.9
0.79
0.91
0.96
0.51
0.63
0.54
0.99
0.73
0.3
E
DisAp-mCherry immuno-EM serial sections

Figure S1. **Identification of the KF component DisAp and its phylogenetic conservation.** (A) A schematic of the workflow used to identify the *disA-1* locus. (B) The sequence for *disA-1* cDNA highlighting the introduction of a premature stop codon within intron 1. (C) The allele frequency (number of reads supporting the nonreference base) of the 206 candidate mutations in the *disA-1* mutant pool sorted in increasing order. Only 26 of these candidates fit the recessive nature of the *disA-1* mutation. (D) Phylogenetic analysis of the *DISA* gene family. A Bayesian-based phylogenetic tree was constructed using Mr. Bayes v3.2.2. The resulting 50% majority rule consensus tree is shown with percentage posterior probabilities indicated at each node. Highly duplicated genes from the reference genome of *Paramecium tetraurelia* are collapsed with the number of closely related orthologues indicated. Tt, *Tetrahymena thermophila*; Im, *Ichthyophthirius multifiliis*; Pt, *Paramecium tetraurelia*; Ot, *Oxytricha trifallax*; Gi, *Giardia intestinalis*; Gl, *Giardia lamblia*; Nc, *Neospora caninum*; Tg, *Toxoplasma gondii*; Pf, *Plasmodium falciparum*. (E) Serial sections from the DisAp-mCherry immuno-EM used in Fig. 1 G.

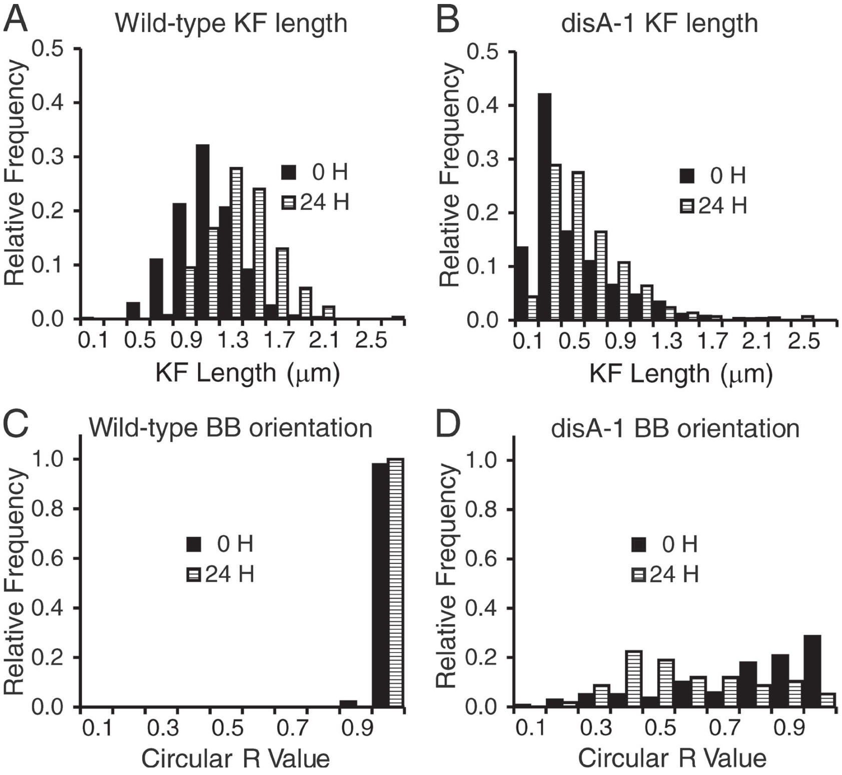


Figure S2. **WT and disA-1 phenotype distribution.** (A and B) Relative histograms for KF length for 0 h and 24 h temperature shifted WT (A) and disA-1 cells (B). The averaged data are in Fig. 2 D. (C and D) Relative histograms for BB orientation for 0 h and 24 h temperature-shifted WT (C) and disA-1 cells (D). The averaged data are in Fig. 2 F.

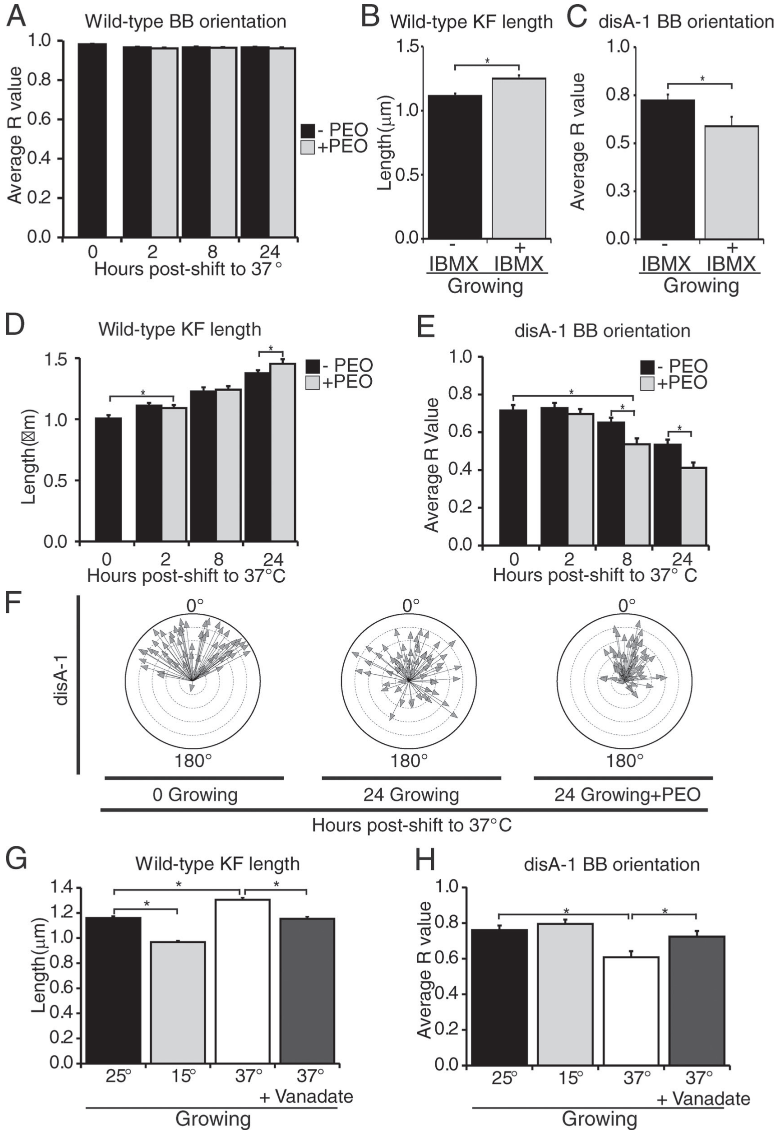


Figure S3. **KFs and the disA-1 phenotype are modulated by ciliary beating.** (A) WT BB orientation is not affected by high-viscosity media. $n > 100$. (B and C) The cAMP agonist IBMX lengthens WT KFs (B; $n > 141$ KFs) and increases disA-1 BB disorientation (C; $n = 30$). (D) High-viscosity media exacerbates temperature-induced KF elongation. Mean KF length for WT cells grown at 37°C in SPP without PEO (black) or SPP with PEO (gray) is shown. $n > 298$ KFs. (E) High-viscosity media exacerbates temperature-induced BB disorientation in disA-1 cells. The mean circular R value for disA-1 cells grown is given as in D. $n > 56$. (F) Polar plots of the mean angular measurement for disA-1 KFs in high-viscosity media. $n > 56$. (G) WT KF length is reduced at low temperature (15°C), and temperature-induced KF elongation is abolished by the dynein inhibitor sodium orthovanadate. $n > 259$ KFs. (H) Temperature-induced disA-1 BB disorientation is rescued by the dynein inhibitory sodium orthovanadate. $n = 50$. Brackets indicate the samples being compared and asterisks indicate statistical significance ($P < 0.01$).

Table S1. **TtDISA1 clade members**

Gene identifier	Accession No.	Description	Abbreviation
296004794	XP_002808749	Conserved *Plasmodium* protein, unknown function (*Plasmodium falciparum* 3D7)	Pf DISA
401399829	XP_003880645	conserved hypothetical protein (*Neospora caninum* Liverpool)	Nc DISA
237841033	XP_002369814	hypothetical protein, conserved (*Toxoplasma gondii* ME49)	Tg DISA
403369035	EJY84356	hypothetical protein OXYTRI_17902 (*Oxytricha trifallax*)	Ot DISA5
308163082	EFO65444	Hypothetical protein GLP15_2488 (*Giardia lamblia* P15)	Gl DISA1a
253746803	EET01832	Hypothetical protein GL50581_905 (*Giardia intestinalis*	Gi DISA1
159115484	XP_001707965	Hypothetical protein GL50803_14341 (*Giardia lamblia* ATCC 50803)	Gi DISA1b
229595185	XP_001019366	Hypothetical protein TTHERM_00388620 (*Tetrahymena thermophila*)	Tt DISA4
145523770	XP_001447718	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA3a
145502961	XP_001437458	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA3b
118354193	XP_001010359	IQ calmodulin-binding motif family protein (*Tetrahymena thermophila*)	Tt DISA3
471235683	XP_004039726	IQ calmodulin-binding motif family protein, putative (*Ichthyophthirius multifiliis*)	Im DISA2
146176274	XP_001019897	Hypothetical protein TTHERM_00588900 (*Tetrahymena thermophila*)	Tt DISA2
145527424	XP_001449512	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2a
145497615	XP_001434796	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2b
145475419	XP_001423732	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2c
145533184	XP_001452342	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2d
145545841	XP_001458604	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2e
145550195	XP_001460776	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2f
145517692	XP_001444729	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2g
145534684	XP_001453086	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2h
145529105	XP_001450341	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2i
145500718	XP_001436342	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2j
145517396	XP_001444581	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA2k
403367908	EJY83781	Hypothetical protein OXYTRI_18485 (*Oxytricha trifallax*)	Ot DISA2a
403351290	EJY75132	Hypothetical protein OXYTRI_03485 (*Oxytricha trifallax*)	Ot DISA2b
146183721	XP_001026900	Hypothetical protein TTHERM_00941400 (*Tetrahymena thermophila*)	Tt DISA
471229666	XP_004035676	Hypothetical protein IMG5_092570 (*Ichthyophthirius multifiliis*)	Im DISA1
145503081	XP_001437518	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1a
145523660	XP_001447663	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1b
145518359	XP_001445057	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1c
145506455	XP_001439188	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1d
145530515	XP_001451035	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1e
145502176	XP_001437067	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1f
145523051	XP_001447364	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1g
145537045	XP_001454239	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1h
145519858	XP_001445790	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1i
145503562	XP_001437756	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1j
145553443	XP_001462396	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1k
145476379	XP_001424212	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1l
145525064	XP_001448354	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1m
145538085	XP_001454748	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1n
145542877	XP_001457125	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1o
145544358	XP_001457864	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1p
145513414	XP_001442618	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1q
145533745	XP_001452617	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA1r
403331043	EJY64442	Hypothetical protein OXYTRI_15526 (*Oxytricha trifallax*)	Ot DISA4a
403339270	EJY68893	Hypothetical protein OXYTRI_10490 (*Oxytricha trifallax*)	Ot DISA4b
145487977	XP_001429993	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4a
145491824	XP_001431911	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4b
145515856	XP_001443822	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4c
145493704	XP_001432847	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4d
145476945	XP_001424495	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4e
74830847	CAI39115	KdD6 (*Paramecium tetraurelia*)	Pt KdD6
145538552	XP_001454976	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4g
145502275	XP_001437116	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4h
145524201	XP_001447928	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4i
118366717	XP_001016574	Hypothetical protein TTHERM_00188980 (*Tetrahymena thermophila*)	Tt DISA5

Table S1. **TtDISA1 clade members** *(Continued)*

Gene identifier	Accession No.	Description	Abbreviation
471223386	XP_004030473	Hypothetical protein IMG5_160360 (*Ichthyophthirius multifiliis*)	Im BBC39a
471221309	XP_004027596	Hypothetical protein IMG5_180510 (*Ichthyophthirius multifiliis*)	Im BBC39b
118387667	XP_001026936	Hypothetical protein TTHERM_00688340 (*Tetrahymena thermophila*)	Tt BBC39
145492272	XP_001432134	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4j
145499182	XP_001435577	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4k
145492266	XP_001432131	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt DISA4l
145479681	XP_001425863	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29a
145539800	XP_001455590	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29b
145513234	XP_001442528	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29c
145529630	XP_001450598	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29d
145521063	XP_001446387	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29e
145549333	XP_001460346	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29f
145483303	XP_001427674	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29g
145537508	XP_001454465	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29h
145541010	XP_001456194	Hypothetical protein (*Paramecium tetraurelia* strain d4-2)	Pt BBC29i
118387711	XP_001026958	Hypothetical protein TTHERM_00688560 (*Tetrahymena thermophila*)	Tt BBC29
471220004	XP_004025084	Hypothetical protein IMG5_192870 (*Ichthyophthirius multifiliis*)	Im BBC29

CHAPTER 6

WHAM: IDENTIFYING STRUCTURAL VARIANTS OF BIOLOGICAL CONSEQUENCE

6.1 Abstract

Existing methods for identifying structural variants (SVs) from short-read datasets are inaccurate. This complicates disease-gene identification and efforts to understand the consequences of genetic variation. In response, we have created WHAM (Whole-genome Alignment Metrics) to provide a single, integrated framework for both structural variant calling and association testing, thereby bypassing many of the difficulties that currently frustrate attempts to employ SVs in association testing. Here we describe WHAM, benchmark it against two other widely used SV identification tools, Lumpy and Delly, and demonstrate WHAM's ability to identify and associate SVs with phenotypes using data from humans, domestic pigeons, and vaccinia virus.

6.2 Author summary

The ability to rapidly and accurately identify structural variants associated with diseases or other phenotypic variations is an ongoing bioinformatics challenge. WHAM is the first tool that provides means to both identify structural variants and to associate them with phenotypes. WHAM's structural variant detection and genotyping algorithms have been designed specifically for association testing with target and background individuals. WHAM maintains high sensitivity, ensuring it does not miss putative causal variants. In its association-testing mode, WHAM can filter out false positives that are shared between the cases and controls. WHAM works on both sequenced individuals and pooled data, and its association framework can be used in conjunction with a variety of SV detection tools. WHAM was designed for ease of use and has clear documentation. WHAM analyses can easily be performed in parallel to standard variant calling pipelines. We provide two examples of how WHAM can be used to find structural variants associated with traits of interest. These examples show that WHAM is a flexible and reliable tool for genotype-phenotype association testing and disease gene discovery.

6.3 Introduction

Structural variation (SV) is a major source of phenotypic variation (Axelsson et al., 2013; Chan et al., 2010; Gemayel et al., 2010; Perry et al., 2008) and human disease (McCarroll and Altshuler, 2007; Stankiewicz and Lupski, 2010; Weischenfeldt et al., 2013). Unfortunately, detecting SVs in short-read sequence data is challenging (Onishi-Seebacher and Korbel, 2011). Moreover, using SVs in association studies remains problematic, primarily due to three technical difficulties. First, SV callers suffer from both high false positive and false negative rates (McCarroll and Altshuler, 2007). Second, the breakpoints of SVs are highly variable, making it difficult to detect an association between a phenotype and a complex ensemble of overlapping SVs (Kidd et al., 2008). Lastly, to our knowledge, there are no statistical tests specifically designed to identify SV enrichment in cases vs. controls within a framework amenable with high throughput sequence analysis. As we demonstrate, WHAM (**Wh**ole-genome **A**lignment **M**etrics) effectively addresses these problems.

Current mapping-based algorithms (Chen et al., 2009; Hart et al., 2013; Layer et al., 2014; Marschall et al., 2012, 2013; Rausch et al., 2012; Sindi et al., 2012) use various mapping attributes such as read depth (RD), paired-end mapping (PEM), split-read mapping (SRM), and soft-clipping to identify SVs. Tools that incorporate more than one of the short-read mapping signals, like Lumpy, Delly, and GASVPro, show improvements over their predecessors that only use a single attribute to discover SVs (Hart et al., 2013; Layer et al., 2014; Rausch et al., 2012; Sindi et al., 2012). SV callers have varying accuracy for different classes of SVs, and some have specifically designed heuristics for the identification of certain SV types (e.g., Delly). Because of this, ensemble methods, such as iSVP, SVmerge, and bcbbio-nextgen, have emerged. These methods integrate SV calls from multiple tools to improve accuracy (Chapman; Mimori et al., 2013; Wong et al., 2010).

Other approaches for identifying structural variants use sequence assembly methods in order to pinpoint SVs. There are two main assembly-based methods for SV detection: *de novo* and local. *De novo* assembly can identify SVs with great accuracy (Li et al., 2011), but also can be prohibitively expensive in computational terms. There are also postprocessing barriers for examining SVs from multiple individuals using *de novo* assembly. For example, synchronizing the coordinates of SVs present from *de novo* assemblies across many individuals is not a trivial task. Multiple sequence alignments provide one approach, but this is computationally expensive and is itself subject to systematic errors (Kemena and Notredame, 2009). Another option for assembly-based SV detection is local assembly. This approach uses read mapping information to confine assembly to putative breakpoints within a genomic range, thus circumventing the need for whole genome assembly (Chen et al., 2014; Narzisi et al., 2014; Quinlan and Clark, 2010). One drawback of local assembly is that it cannot discover sequences of large novel insertions, which might only be revealed by *de novo* assembly—and alignment of reads to a reference genome remains problematic. Lastly, gains made possible by local and *de novo* assembly are dependent upon higher read depths. Given finite resources, sequencing fewer individuals at a higher depth compromises power for conducting downstream association testing (Kim et al., 2010; Sims et al., 2014).

Whereas other modern SV callers strike a balance between sensitivity and specificity (Alkan et al., 2011), WHAM is configured to err on the side of sensitivity. This is because using structural variant calls for genotype-phenotype association discovery benefits from high-sensitivity and high-quality genotype calls. Under calling variants can result in the failure to detect biologically relevant SVs. Furthermore, specificity is less of an issue for association testing; so long as false positives occur randomly or systemically across cases and controls, they will fall by the wayside during association testing. WHAM's highly sensitive SV identification and genotyping algorithms are tuned for association testing. As we show, WHAM is able to pinpoint SVs in pooled and genotypic data that cause phenotypic variation. WHAM thus fills the need for a fast, easy to use, and highly sensitive SV caller and association-testing tool that is compatible with most standard variant calling pipelines.

6.4 Results and Discussion

WHAM integrates multiple mapping-based signals to identify putative SV breakpoints. Both individual genome and populations of individuals (pooled sequencing) datasets can be processed with WHAM. Additionally, if two cohorts of genomes are provided (target and background), WHAM can be used to conduct an association test. This provides means both to identify SV's with genotype phenotype associations and to filter SV false positives. WHAM also provides means for classification of SV type (deletions, duplications, inter-chromosomal events/insertions, and inversions). Classification is performed *post hoc*, as WHAM conducts genotyping and association testing independent of the SV type. Here we explore the accuracy of WHAM's SV detection and genotyping by first using simulated short-read datasets, followed by two whole genome human datasets. We also use WHAM to identify biologically important structural variants in nonhuman data.

6.4.1 Validation of WHAM using simulated data

We first examined the performance of WHAM's SV detection heuristic and compared it to two other SV callers, Delly (Rausch et al., 2012) and Lumpy (Layer et al., 2014), using simulated whole genome sequencing (WGS) data. Synthetic reads were generated for 10x and 50x whole genome coverage with simulated occurrences of four classes of structural variants (deletions, duplications, interchromosomal events/insertions, and inversions; see Materials and Methods for details). Simulated insertion events were created by placing sequences from other chromosomes into alternate locations mimicking inter-chromosomal copy number variants; we will refer to these events as insertions throughout the rest of this section. We chose to benchmark Delly and Lumpy because both tools can identify multiple types of SVs, are widely used, and are easy to install and run directly from BAM files. Lumpy also provides a point of reference against GASVPro and Pindel, as it has already been benchmarked against these tools under matched simulation conditions (Sindi et al., 2012; Ye et al., 2009). We used a previously published interval size (regions defined by 25 bp up and downstream of each simulated variant breakpoint) as "truth intervals" to determine true positive calls, unless otherwise noted (Layer et al., 2014). A SV is considered a true positive only if both of the called breakpoints lie within a "truth interval." For specific details regarding the simulations, see Materials and

Methods.

WHAM, Lumpy, and Delly were run in their default modes across the simulated data to identify SV breakpoints. For a high depth simulated dataset (50x), WHAM and Lumpy have comparable sensitivity overall (0.94 and 0.90, respectively), while Delly has slightly lower sensitivity (0.84). The structural variant size drives the largest differences in sensitivity between the three tools (Fig. 6.1A). For example, neither Lumpy nor Delly is able to detect many of the smaller duplications (50 bp and 100 bp; collapsed into the 0.05-1kbps interval [Fig. 6.1A]). Delly's limitation in detecting smaller SVs (~300bp) has been acknowledged by the authors. For smaller SVs (<60bp) the sensitivity of WHAM is generally 2-3 times greater than the other tools (Figure 6.1A; 0.05-1kbps interval). All three tools have similar sensitivity for detecting simulated SVs greater than 1kb in size with Lumpy showing slightly higher sensitivity. Given that the observed frequency of SVs follow a power law distribution with respect to size, we would expect that Delly and Lumpy will miss many more SV calls than WHAM on real biological datasets (Abecasis et al., 2010; Pang et al., 2010). Compared to its performance on other classes, WHAM has the lowest sensitivity for insertions in a 10x coverage simulated dataset (Fig. 6.1A). This is due to WHAM incorrectly identifying one of the two breakpoints at lower depths (data not shown), which ceases to be a limitation at higher depth of coverage. These sensitivity assays demonstrate that WHAM excels at finding small SVs (less than 1kb) while maintaining similar performance to Lumpy and Delly for SVs greater than 1kb.

Next we assayed the false discovery rate (FDR) of the three tools on the same simulated data. WHAM has an FDR of 0.05 at 50x with 50bp of slop, which is higher than Delly (0.02), but lower than Lumpy (0.11) (Fig. 6.1B). Lumpy has the lowest overall FDR if insertions are excluded. Reducing the amount of slop added to the confidence intervals slightly increases the FDR for Delly and Lumpy, but not WHAM. WHAM's FDR can be attributed to misclassification of SV type and failures when identifying both breakpoints. All three tools exhibited a positive correlation between depth and FDR when comparing the 10x and 50x datasets. For example, Delly's FDR for deletions nearly doubles in the 50x relative to the 10x data. All three tools had elevated FDRs for insertion events. This is because our simulated insertions create interchromosomal duplications, which increase mapping errors, leading to false positive SV calls. As expected, the FDRs for the simulated data are much lower than the human benchmarks, presented later.

To assess the breakpoint accuracy of the tools, we removed the confidence intervals for deletions and then incrementally added 1bp-500bp of bi-directional slop to both breakpoints (Fig. 6.1C). WHAM has the highest positional accuracy for deletions of the three tools, as it has the highest sensitivity (0.75) with only 1bp of slop. Lumpy exhibits a marked gain in sensitivity from 10bp to 25bp of slop as it is designed to detect "soft" breakpoint boundaries (Layer et al., 2014), whereas Delly's sensitivity exhibits an increase from 1bp to 5bp of slop. In contrast, WHAM maintains a near constant sensitivity down to 5bp of slop, after which WHAM's sensitivity drops, but remains greater than 0.75. WHAM maintains sensitivity at small intervals by relying on highly accurate mapping and soft clipping. WHAM, unlike the other tools, uses soft-clipping information to call small SVs. WHAM's breakpoint sensitivity is important for maintaining power during association testing. The power to detect an association between a SV and a disease is diminished when breakpoints are miscalled within a cohort of affected individuals. All three tools showed improved breakpoint detection at higher depth (Fig. 6.1C). The high positional sensitivity shows that mapping-based methods can reliably localize SV breakpoints down to a 3bp interval. This small interval provides sufficient accuracy for association testing.

Genotyping accuracy is critical for association testing; therefore, we wanted to assay WHAM's genotyping accuracy. WHAM provides bi-alleleic genotype calls (homozygous, heterozygous, homozygous nonreference) for individual(s). The proportion of correctly called genotypes is reported in Fig. 6.1D for Delly and WHAM (Lumpy does not directly provide a genotype call). For this benchmark, we used the same simulated dataset used in Fig. 6.1A-C, in which all simulated structural variants were homozygous nonreference. True positives were defined as structural variant calls genotyped as homozygous nonreference. Delly exhibited slightly higher genotype sensitivity for deletions, whereas WHAM has higher sensitivity for inversions (Fig. 6.1D). Interestingly, both WHAM and Delly fail to correctly genotype duplications of any size (Fig. 6.1D); all simulated homozygous duplications were genotyped as heterozygous. The catastrophic genotyping failure rate for duplications can be attributed to mapping artifacts. Duplications create nonunique sequences and mapping software commonly deals with this problem by random assignment of reads between breakpoint boundaries, making the assessment of ploidy extremely difficult. Both WHAM and Delly also had marginal sensitivity for insertions as they were simulated as interchromosomal duplications.

Collectively, these simulations show that WHAM provides a robust means for SV identification and genotyping. Compared to the other two tools, WHAM excels at finding smaller structural variants across all simulated SV classes and has the highest breakpoint sensitivity. This is significant as SVs are distributed geometrically with respect to length, thus shorter SVs comprise the overriding majority of real events (Abecasis et al., 2012; Lappalainen and Lopez, 2013). As we will show later, WHAM also maintains high sensitivity for real human short-read data, but at the cost of a much higher false discovery rate.

6.4.2 Validation using human WGS data

We began our human benchmarks with NA12878, the best characterized genome. For the truth set, we used the One Thousand Genomes SV dataset (phase III submitted calls)(Abecasis et al., 2012) downloaded from dbVar (estd214)(Lappalainen and Lopez, 2013). This dataset contains 2,975 integrated SV calls, ranging in size from ~200bp to ~900Kbp. Deletions, large indels, and mobile element insertions make up the majority of the NA12878 subset. It is worth noting that Delly calls are represented in this truth set, but not WHAM or Lumpy calls. We ran each SV caller, as per best practices, over NA12878, generating between ~10K and 600K SV calls (Table 6.1). The expected number of SV calls in NA12878 depends on the SV size range and the tolerated FDR. The One Thousand Genomes project imposes at 10% FDR for structural variants, improving the accuracy at the expense of sensitivity (Abecasis et al., 2012). Therefore, the One Thousand Genome Project SV calls for NA12878, while highly accurate, are very much an underestimation of the total number of SVs for NA12878. In an upward extreme, another group (Bickhart et al.) reported over a million deletions in NA12878 (Bickhart et al., 2015). The high number of SV calls for NA12878 made by WHAM, reported in Table 6.1 and reported by Bickhart et al., highlights the importance of *post hoc* filtering prior to downstream analyses. Table 6.1 lists the three filters we used to improve both the sensitivity and FDR of the tools we benchmarked. The first filter removes SV calls where either breakpoint was located in a low complexity region (Li, 2014). The second filter removes sites where the depth of coverage was drastically higher than average. The last filter removes breakpoints that overlap NIST NA12878 indels (Zook et al., 2014). Removing indels lowers WHAM's FDR as small indels are not found in the Phase III SV call set (they are present in the integrated SNV call set). After all filters were combined, WHAM, Delly, and Lumpy had 250.7K, 40.6K and 3.7K SV calls, respectively (Table 6.1). These steps removed several hundred thousand WHAM calls. However, recall that Delly was employed to produce the Phase III NA12878 'truth' set used here. We confined our subsequent benchmarks to the two most common classes of SVs in the Phase III dataset, deletions (976), and Mobile element insertion (1,070; MEI).

The three SV callers have similar performance for deletions (Fig 6.1A);

Lumpy has the highest sensitivity (0.63) and lowest FDR (0.75) overall. The sensitivity for all three tools starts at ~0.75 in the 150bp-1Kbp interval and tapers off to ~0.25 for SVs greater than 10Kb (Fig. 6.2A). The FDR was high for the smallest and largest size categories (100bp-1Kbp and 10Kb-1Mb). Between these two categories, the smaller size window contained the largest number of false positives and true positives for each tool. For example, Delly has over 167 true positives and 28,769 false positives in the smallest size category. WHAM's sensitivity (0.61) was lower than Lumpy (0.63), but not Delly (0.60) and overall, WHAM's FDR (0.80) was below Delly's (0.98). We next assayed the three tools' ability to detect mobile element insertions (MEIs). There are more MEIs (1,070) in the NA12878 truth set than deletions (976). We intersected all Delly and Lumpy calls, regardless of classification with the NA12878 MEIs. In contrast, because WHAM has the ability to classify insertions, we only intersected the WHAM insertions with the MEI calls. Unlike the previous benchmarks, if either breakpoint of a putative MEI overlaps a Phase III MEI, it is classified as a true positive. Novel insertions and MEIs only have one breakpoint relative to the reference sequence. Out of the three tools, WHAM was the only SV caller that detected a sizable number of MEIs albeit with a high FDR (Fig 6.1B). If all SV classes are grouped together, WHAM has the highest sensitivity (0.70) and the highest FDR (0.98) for the NA12878 dataset. The Phase III NA12878 SV truth set is highly conservative; therefore, we expect that as more real SV calls are integrated into the One Thousand Genomes Project, the sensitivity and FDR of each tool will improve.

For a second, independent, human benchmarking experiment, we used the recently published single-molecule, real-time (SMRT) sequencing dataset of a hydatidiform mole cell line (CHM1) (Chaisson et al., 2014). Both PacBio and Illumina sequence data were generated from DNA recovered from CHM1 cells. The 101 bp Illumina reads and PacBio (~8kb average length) reads cover the haploid genome to 40.7x and 36.6x depth, respectively. The structural variant calls from the PacBio single molecule sequencing were generated by first identifying putative SV breakpoints followed by local assembly (see supplement of (Chaisson et al., 2014)). The PacBio SMRT SV calls are a good standard for validating WHAM's performance on the related Illumina datasets because PacBio SMRT sequencing does not require DNA cloning or amplification, two common sources of sequencing artifacts. Moreover, the absence of allelic heterogeneity in the haploid CHM1 genome facilitates accurate assembly (Chaisson et al., 2014; Steinberg et al., 2014). Additionally, PacBio reads can capture small and moderately sized SVs internally within a read, providing a more accurate source for detecting SVs. We analyzed the Illumina data with WHAM, Delly, and Lumpy, and compared their SV calls to the 11,311 SMRT deletions and 15,330 insertions (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation).

WHAM ses Lumpy and Delly in sensitivity for the CHM1 dataset. WHAM has the highest number of true positive deletion calls (1,983), followed by Lumpy (1,759) and Delly (1,547) (Fig. 6.2C; both size intervals combined). This finding is in contrast with the NA12878 benchmarking where WHAM had the lowest sensitivity for deletions. We attribute this difference to the size distributions of the two truth sets. The CHM1 dataset is enriched for smaller SVs whereas the Phase III NA12878 calls focus on larger SVs (greater than 200bp). WHAM has increased sensitivity for smaller structural variants (Fig. 6.2C; 25bp-500bp). In the larger size category (500bp-100k), Lumpy has a higher sensitivity than WHAM, but not Delly. WHAM's false discovery rate (0.47) was higher than Delly's (0.41) and Lumpy's (0.34) (Fig. 6.2C; both size intervals combined). Next we turned our attention to the CHM1 insertions. WHAM had the highest sensitivity, finding over 38% of the CHM1 insertions (Fig. 6.2D). WHAM's exceptionally high FDR for insertions (0.86) is a result of not filtering. We did not apply filtering as 9,445 of the 15,373 CHM1 insertions overlap low complexity regions of the genome (Chaisson et al., 2014; Li, 2014). Lastly, we examined the distribution of deletions sizes called by WHAM, Delly, and Lumpy compared to the CHM1 dataset (Fig. 6.2E). WHAM's size distribution for deletions closely tracks the CHM1 dataset. Both datasets are enriched for deletions less than 100bp, which is concordant with previous studies (Abecasis et al., 2012; Mills et al., 2006). The peaks in the size distributions at 300bp and 6000bp correspond to ALUs, STRs, and LINE-1 elements. The variability between the size distribution of the tools suggests that each tools is well suited for a slightly different size class.

The results from benchmarking on real data have several important implications. First, WHAM is a sensitive structural variant caller and achieves comparable, or better, performance relative to other commonly used structural variant callers, but at the cost of higher FDR. Importantly, WHAM provides robust means for discovering small structural variants and mobile element insertions. MEIs comprise 55% of all NA12878 SV calls in the One Thousand Genomes Project (phase III). Second, the low overlap between SV classes among the tools tested here supports the power of integrated SV call sets. Frameworks, like the approaches of bcbio, which acts by combining SV calls from a variety of callers (including WHAM), can capture a greater swath of genetic diversity while also providing higher confidence for concordant allele calls across varying heuristic methods (Chapman; Mimori et al., 2013; Wong et al., 2010).

6.4.3 Identifying candidate SVs with WHAM's association test

Although all three tools have relatively similar sensitivity for NA12878 deletions, any critical appraisal of their performance must take into account the very high false discovery rates of all three tools. Using the NA12878 deletion data, for instance, WHAM's FDR is 0.80, Delly's is 0.98, and Lumpy's is 0.75. These values illustrate just how difficult SV discovery is using short-read data. However, for purposes of genotype-phenotype association, high false discovery rates are tolerable so long as false positives are either randomly distributed across cases and controls (nondifferential error), or else systematic, e.g., called in every individual. In both scenarios, false positives will cancel out in an association test. Thus given a reasonable true positive rate, robust association signals will be obtained even in the face of very high FDR. Bearing these issues in mind, in the analyses below, we demonstrate the suitability of WHAM for association testing and demonstrate the efficacy of its association test using real biological datasets from domestic pigeons and vaccinia virus.

We first sought to test if WHAM's nondifferential false discovery rate creates spurious signals of association. To examine this, we used a cohort of individuals with a high degree of genetic relatedness such that if they were assigned randomly into two groups for association testing, there should be little to no differentiation. We chose the CEPH/Utah Pedigree 1463, comprised of seventeen individuals across three generations (Abecasis et al., 2012; Illumina). This pedigree should not harbor appreciable levels of population stratification, thus removing a potential confounding source of false positive associations in our sampling. WHAM was run in default mode three times, randomly dividing the pedigree into two groups of eight individuals for assignment to either target or background groups. One genome was excluded each round so that the target and background had the same number of individuals. True and false SV calls were assigned according to their proximity to the phase III One Thousand Genomes Project SV calls using a 50 bp truth interval. In total, 16,470 association tests were run for the true positive SV calls, while 380,005 were run for the false positives. Comparing the distributions of WHAM's LRT p-values (Chi-sq one degree of freedom) between the groups showed a significant difference between the true and false SVs, shown in Fig. 6.3 (Two-sample Kolmogorov–Smirnov [KS] test D = 0.0948, p-value = 2.2e-16). The median p-value for the true positive group was 0.66 and 0.69 for the false positive group (1.03 times higher). To see if the significant difference is robust to the number of variants assayed, we subsampled 100 WHAM p-values for both groups and the

KS test was rerun. Over 1000 iterations, only 362 of the 1000 KS tests achieved significance. This suggests that there is a small, albeit significant, difference between true positives and false positives. Together, this demonstrates that WHAM's false SV calls are only expected to slightly inflate the number of spurious associations.

While WHAM has a high false discovery rate for SV detection, WHAM's association-testing framework is robust to many of these errors as they are non-differential between the cases and controls. To demonstrate that WHAM's high FDR for SVs does not hinder association studies, we used WHAM on both genotypic (pigeon) and pooled (viral) datasets. In the pigeon dataset, WHAM was used to remap a SV causative for the recessive red pigmentation trait and in the viral dataset, we show that WHAM reliably identifies the breakpoints of a duplication involved in viral adaptation.

6.4.4 Identifying the genetic basis of recessive red coloration in domestic pigeons

Pigeon fanciers have selected for a wide range of phenotypic variation in domestic pigeons over thousands of years. These traits include plumage patterns, behavior, size, and pigmentation (Shapiro et al., 2013). Several alleles in three genes, *Tryp1*, *Sox10*, and *Slc45a2,* were recently identified that affect the melanin synthesis pathway (Domyan et al., 2014). For example, birds homozygous for a 7-kb deletion spanning a melanocyte-specific enhancer of *Sox10* (*e1* allele) have reduced expression of *Sox10* and its target *Tyrp1*, resulting in the ‘recessive red’ color phenotype. Using a previously generated WGS dataset, we examined the power of WHAM’s association test to re-identify the *e1* allele. In conjunction with the WHAM analyses, we also ran the same association test for SNPs (implemented in pFst Chapter 2) and Delly SV genotype calls. Lumpy was excluded from these analyses, as it does not provide a method for joint genotyping, which is required for WHAM’s association framework.

WHAM re-identified the *e1* allele as the best genome-wide candidate for recessive red using a likelihood ratio test (LRT; Fig. 6.4A, Fig. 6.4C). The LRT implemented in WHAM measures the differences in allele frequencies based on the genotype calls at every SV position in the genome. Five recessive red and six wild type birds were processed with WHAM to identify SVs and conduct association testing. The highest WHAM LRT scores were present on scaffold974 at the two PCR confirmed breakpoints of the *e1* allele (Fig. 6.4C). Because the pigeon reference genome was assembled from a recessive red bird that harbored the *e1* deletion allele, WHAM indirectly identified the location of the deletion by identifying an “insertion” in the wild-type birds, relative to the reference genome. Delly was unable to identify this allele, because it was not designed to identify novel insertions (Fig. 6.4C). WHAM also detected several inversions that are contained within the *e1* allele and close to the deletion breakpoints. Delly additionally failed to detect these inversions. The increased LRT scores (converted to p-values) around the *e1* allele are attributed to linkage disequilibrium since *e1* is on a shared haplotype. This linkage is more notable in the SNP data, which has a much higher density of variants. The p-values from WHAM's association test fit a uniform distribution, suggesting little to no population stratification between the cases and controls in domestic pigeon (Fig. 6.3B). Importantly, WHAM's demonstrated high false discovery rate did not affect our ability to rediscover the *e1* allele. This analysis demonstrates the utility of WHAM for rapidly identifying candidate SVs for simple Mendelian traits.

6.4.5 Identifying adaptive structural variation in vaccinia virus populations

Structural variants in the form of gene copy number variation (CNV) in DNA virus genomes provide a mechanism for rapid virus adaptation to host immune defenses (Brennan et al., 2014; Elde et al., 2012; Erlandson et al., 2014; Slabaugh et al., 1989). For example, frequent recombination events creating tandem gene duplications have been observed in vaccinia virus (VACV) as a means of adaptation to the human antiviral host factor protein kinase R (PKR) during experimental evolution (Elde et al., 2012). In this system, selective pressure was placed on the virus by deleting the E3L gene encoding a strong PKR inhibitor, leaving only a weak PKR inhibitor encoded by the K3L gene (Beattie et al., 1995). Experimental evolution of this ΔE3L virus in HeLa cells revealed that copy number expansion of the K3L gene provides gains in viral fitness. To test whether CNV is a common mechanism of adaptation and determine whether WHAM is an effective tool to detect and characterize such events, we passaged the ΔE3L VACV strain ten times in a different cell line derived from primary human fibroblasts.

We analyzed short-read sequencing data from viral genomes obtained from both a virus population after ten passages and the parental ΔE3L strain for comparison. This analysis revealed areas of structural variation in the adapted viral population. Plotting read depth across genomic positions revealed a spike in read depth corresponding to the K3L locus that is only present in the adapted strain (Fig. 6.5A). This is consistent with previous work in which a similarly large increase in depth corresponded to increased K3L copy number as a means of adaptation (Elde et al., 2012). To determine the exact position of the recombination event generating the CNV, and to find any novel structural variants, we performed SV calling using either WHAM or Lumpy. Using similar filtering schemes, WHAM identified 6 SV calls in the adapted viral population, compared to 20 SV calls identified by Lumpy (Fig. 6.5A). Overlaying SV calls on the read depth plots shows the increased specificity of using WHAM to identify SVs (Fig. 6.5A).

WHAM analysis identified four SVs in the parental strain, and an additional two in the adapted strain. Notably, all six SVs map near the K3L locus or the E3L deletion (Fig. 6.5B, Table 6.2). The two breakpoints near the K3L locus were only identified in the adapted population, suggesting that the SV was not present in the parental strain. These two breakpoints have very high read support, indicative of the same recombination event dominating throughout the adapted viral population (Table 6.2). Indeed, when we specifically amplified and sequenced the region around the K3L breakpoint, we identified a single breakpoint in the adapted strain, but could not detect any SV in the parental strain at this location. Importantly, the WHAM-identified breakpoints match the exact positions of the breakpoint identified by PCR and Sanger sequencing (Fig. 6.4B). Thus WHAM is able to identify SVs in viral populations, down to single nucleotide accuracy. This analysis also suggests that K3L CNV is a common mechanism for VACV to overcome the antiviral PKR defense pathway.

Surprisingly, the other four breakpoints WHAM identified with high read support map near the E3L deletion (Fig. 6.4C, Table 6.S1). This is unexpected because E3L was originally replaced with a β-galactosidase (β-gal) selective marker, creating an insertion much larger than the reads from deep sequencing. However, the ΔE3L virus was originally engineered to express β-gal under the control of the VACV 11K promoter. This promoter naturally drives expression of the F17R gene, and is thus present in the reference genome at the F17R locus approximately 5kb upstream of E3L. Therefore, one end of the E3L deletion is supported by split reads mapping to the natural viral promoter. The other end of the deletion does not have SR support as the β-gal gene itself is not in the reference genome, but WHAM did pick up a breakpoint with mate-pair support at this end. Importantly, the genomic positions identified by direct sequencing of both the parental and adapted strains for each end of the E3L deletion were correctly called by WHAM (Fig. 6.4C). These data show that WHAM can identify both a genetic rearrangement (K3L) and a novel insertion (β-gal) with respect to a reference sequence. In comparison, while Lumpy successfully identified the K3L duplications (also down to the exact base pair on one end), it failed to identify two of the E3L breakpoints detected by WHAM. Overall, three of the five breakpoint positions identified by direct sequencing were called using Lumpy, although the remaining two are in close proximity to one of the called positions. Also, only one of these positions exactly matches the sequenced position, consistent with Lumpy providing a region rather than a specific position. Thus, although this is only one experiment, WHAM shows greater specificity as demonstrated by fewer total SV calls, as well as improved accuracy when compared to Lumpy in analyzing this dataset. While WHAM's low call rate in this example is not consistent with the human data, there are two possible explanations for this trend; the truth sets for the human data are under called, resulting in WHAM's high FDR, or WHAM under calls pooled datasets. The second possibility is unlikely since WHAM correctly identified the breakpoints of all SVs independently verified in the viral dataset (Elde et al., 2012).

Taking a closer look at WHAM calls with mate-pair (MP) support in addition to SR calls, we discovered a complex set of breakpoints around one end of the E3L deletion. For the K3L breakpoint, WHAM only called the two positions of the single breakpoint, whereas it called one additional position with high read support on one end of the E3L deletion (Fig. 6.4B-6.4C). To determine whether these calls represent true variants, we performed PCR and Sanger sequencing across the region spanning from E2L to E4L, which includes the entire β-gal cassette. This analysis revealed SVs in both the parental and adapted strains that contain partial deletions of the β-gal and the 11K promoter. Thus WHAM correctly identified a previously unknown variable deletion (Fig. 6.4C). We have two hypotheses to explain the appearance of variable deletions in this region. First, in the absence of selection on the β-gal marker gene, there is a fitness cost to carrying the engineered marker, so viruses losing this region have a fitness advantage compared to ones retaining it. Alternatively, using a VACV promoter for β-gal expression present at a second location only ~5kb away in the genome might promote localized recombination in this region of the VACV genome. These hypotheses are not mutually exclusive and highlight how genetically engineered virus strains may not always be homogenous.

In addition to identification of SVs from sequencing individual genomes, this analysis demonstrates that WHAM is able to detect variable structural changes within polymorphic populations. This provides an example of WHAM's utility as a tool for accurate detection of SVs in rapidly changing microbial populations. Gene amplification can play a major adaptive role in response to selective pressure in both viral (Brennan et al., 2014; Elde et al., 2012; Erlandson et al., 2014; Slabaugh et al., 1989) and bacterial populations (reviewed in (Andersson and Hughes, 2009; Elliott et al., 2013; Romero and Palacios, 1997; Sandegren and Andersson, 2009)), so it is important to accurately define the adaptive potential of structural variants. Recent advances in whole genome sequencing provide a wealth of genetic information about microbial population dynamics, and WHAM provides a tool to rapidly identify potentially adaptive SVs.

6.5 Conclusions

WHAM is a highly sensitive structural variant caller and association-testing tool. It is flexible enough to work on a broad range of data including pooled and diploid individuals. We show that WHAM's SV detection compares favorably with other popular mapping-based SV calling methods and performs well across a number of SV types in both simulated and real datasets. While WHAM, like other SV calling tools, suffers from high false positive rates, we show that this is unlikely to affect the association statistics from WHAM's association testing. WHAM's ease of use also makes it an ideal package for inclusion with integrated SV callers. By simply running WHAM in its default association-testing mode, we were able to identify the causal SV allele of a recessive trait in pigeons. Similarly, WHAM's accurate breakpoint predictions were able to locate a copy number variant in viral populations relative to a parental strain with very high precision.

Future efforts will focus on expanding WHAM's association test to handle locus and allelic heterogeneity. Currently, WHAM does not have ability to detect genotype-phenotype associations for nonoverlapping structural variants. Integrating structural variant size, phylogenetic conservation, and the burden of SVs in the cases versus controls will increase the power and utility of WHAM's association test.

6.6 Materials and Methods

6.6.1 Identification of breakpoint and genotyping

WHAM integrates mate-pair mapping, split read mapping, soft-clipping, alternative alignment, and consensus sequence-based evidence to predict SV breakpoints with single-nucleotide accuracy. WHAM generates a combined pileup (ensemble of reads covering a position of the genome) for all BAM files provided. Reads from all individuals included in joint calling that are soft or hard clipped are hashed by position to identify shared breakpoints. Positions in the pileup where three or more primary reads share the same breakpoint are interrogated as a putative SV. The soft clipped sequences that overhang the breakpoint are collapsed into a consensus sequence using a multiple sequence alignment (MSA) provided in the seqAn library (Döring et al., 2008). WHAM applies three filters to the consensus sequences. Breakpoints are not reported in cases where consensus sequences are shorter than 10bp or contain more than 50% mismatches in the alignment, as they more likely reflect mapping errors rather than allelic heterogeneity. Overlapping alleles that do not share the same breakpoints are reported as independent records in the VCF file, allowing for allelic heterogeneity. Alleles that share an exact breakpoint, but different sequences, fail the mismatch consensus filter and are discarded.

WHAM uses split-read (SR) alignments, mate-pair (MP) positional information, and alternative alignments to find the other SV breakpoint (the breakpoint not present in the current pileup position). WHAM is unaware of past SV calls; therefore, it outputs an SV call for the 5’ and 3’ breaks independently.

Each split read entry in a BAM file reports the other supplemental alignments in the “SA” tag and alternative alignments are reported in the “XA” tag. WHAM processes the cigar strings of the SA and XA tag to identify shared positions as candidate endpoints of the reported SV. WHAM clusters all the candidate breakpoints and rounds their positions to the nearest tenth base pair. The position with the highest number of read support is reported. If the soft-clipped consensus sequence can be aligned to the putative breakpoint region using Smith and Waterman, the breakpoint is further refined to the location of the consensus sequence alignment. The amount of support for the breakpoint is listed in the “SP” info field.

Translocations and structural variants greater than 1Mb undergo additional filtering. These classes of SVs can be highly deleterious genomic aberrations; therefore, we require them to have additional support. Large intrachromosomal SVs require that the other breakpoint (outside pileup position) has at least two reads supporting the exact breakpoint. This same filter is applied to translocations. Additionally, in the case of translocations, if the split reads in the pileup map to more than three different chromosomes, the SV is discarded. This filter removes many false positive SV calls resulting from interchromosomal mapping errors introduced by repetitive sequences.

Genotyping is accomplished using a bi-allelic likelihood model (Li, 2011; Nielsen et al., 2011). Rather than using base quality at the breakpoint position, we use the mapping quality of the read. Each read that contains the breakpoint, internally or soft clipped, is counted as nonreference. Additionally, reads that are discordantly mapped or show signs of an inversion (same strand mate pair mapping) are also considered to be nonreference for use in genotype calling. During joint calling at least one individual must have three reads supporting the alternative allele. This filter prevents randomly shared start and stop soft clipping across individuals from triggering a nonreference allele call.

For best performance, we recommend using BWA mem (Li, 2013) followed by sorting and duplicate removal (duplicate marking is also supported) of the BAM files. The BWA mem algorithm provides soft clipping and split read annotations. Specifically, the “SA” and “XA” optional fields in the BAM files are heavily utilized by WHAM. Supplementary read alignments (0x800 / split reads) can be marked as secondary with no detrimental effect. Marking or removing duplicates is highly recommended as these duplicates cause false positive SV calls. Other mapping software like Bowtie2 (Langmead and Salzberg, 2012) also provides soft clipping, which is sufficient to run WHAM, but produces results with lower sensitivity (data not shown). WHAM can be run on single-end sequencing data, but for best results, paired-end data is recommended.

6.6.2 Classification of SV type

WHAM classifies the type of structural variant by using an ensemble of decision trees (random forest) implemented in scikit-learn (Pedregosa et al., 2011). This approach is similar to another SV caller, forestSV (Michaelson and Sebat, 2012). WHAM’s raw breakpoint calls (in VCF format) are postprocessed by ‘classify_WHAM_vcf.py’ to add SV type to the INFO field. The WHAM classifier provides the SV type in the “WC” info field and probability of each type in the “WP” info field. We use fourteen attributes of a genomic position for the classifier (Table 6.3). Each attribute is a fractional measure reflecting the number of reads that belong to each attribute, normalized the by the read depth at the pileup position. Some of the fourteen attributes have low to no importance for training the model, but we chose to maintain them as they allow further downstream development. The training dataset is derived from our simulated dataset, which includes deletions, insertions/translocations, duplications, and inversions. The k-fold cross-validation implemented in scikit-learn reports a validations rate of ~0.94 for the simulated dataset. A user may create their own training set consisting of a truth set of variants, supplying as many variant types as they see fit. To do this, WHAM should be run over a BAM file containing SVs that have been validated. Then the “AT” info field should be split into a tab-delimited file with the last column providing the validated SV type. The resulting training file should match the format of the file distributed with WHAM. Additionally, WHAM can be extended to annotate as many features as the user sees fit. False positive WHAM SV calls can be annotated and added to a training set. This flexibility makes WHAM extendable to identify many patterns in a pileup that differentiate between SV types.

6.6.3 WHAM's association test

When the "target" and "background" options are enabled, WHAM quantifies the difference between the target and background allele frequencies using a likelihood ratio test (LRT) under a binomial likelihood model with one degree of freedom. The basic LRT used within WHAM has been widely adopted for association studies (Kim et al., 2010; Li, 2011; Yandell et al., 2011). WHAM's LRT has also been implemented in GPAT++, a population genetics library (Kronenberg).

The null model of WHAM's LRT assumes that the allele frequencies of both the target (AF_T) and background (AF_B) groups are the same, while the alternative hypothesis is that the allele frequencies of the two groups come from two separate distributions. The allelic counts in the model come from the genotype calls.

$$D = -2 * ln(\frac{B(N_C,K_C,AF_C)}{B(\ N_T,K_T,AF_T\ ) \times B(N_B,K_B,AF_B)}) \tag{6.6.3.1}$$

The binomial density function (B*(n, k, p)*) is parameterized by the number of successes *n*, the number of trials *k*, and the probability of success *p*. In the current application, *n* is the number of nonreference alleles in the target (N_T), background (N_B), and the target/background combined (N_C). The parameter *k* is the number of alleles in the target (K_T), background (K_B), and the target/background combined (K_C). The probability of success *p*, corresponds to the target (AF_T), background (AF_B), and combined (AF_C) allele frequencies. WHAM reports the *D* statistic in the "LRT" info field. Larger LRT values can indicate that the null hypothesis should be rejected under the assumptions of the binomial model. A chi-sq lookup, with 1 df, can be used to convert the *D* statistic into a p-value.

6.6.4 Simulations and human benchmarks

Nonoverlapping homozygous deletions, duplications, inversions, and translocations/insertions were independently injected into the human reference genome (hg19, GATK resource bundle (DePristo et al., 2011; McKenna et al., 2010)) using SVsim (Faust). SVsim generates insertions by placing fragments from another chromosome into the target site; therefore, we denote insertions as translocations/insertions. Two SV sizes, one and five, were incremented by powers of 10 (1-6) generating SVs from 50bp – 1Mb. From the mutated sequences, 150bp paired-end reads were simulated using DWGsim at two average depths, ten and fifty (Homer). DWGsim added single nucleotide polymorphisms, but no additional SVs. The simulated reads were aligned to the human reference genome using BWA mem in the default mode (Li, 2013). The alignments were converted to BAM files, sorted, and duplicates were removed using Samtools (Li et al., 2009). Each sample was run independently with WHAM, Delly, and Lumpy. WHAM and Delly VCFs were converted to BEDPE format using two scripts distributed with WHAM (WHAMToBedPe.pl and dellyToBedPe.pl). For benchmarking, we used bedtools set operations to determine true and false positives (Quinlan and Hall, 2010). The ‘pairToPair –type both –slop 50 –is’ command was used to find true positives, requiring both putative breakpoints overlap with a single simulated SV. False positives were counted with ‘pairToPair –type notboth –slop 50 –is’. This benchmarking scheme has previously been used and honors the confidence intervals provided by all three tools.

All of Illumina’s platinum genomes (17 member CEPH pedigree 1463) were aligned using BWA mem, sorted with Sambamba (v0.5.0-dev), and duplicates were removed with Samblaster (0.1.20) (Faust and Hall, 2014; Tarasov). The truth set (phase III One Thousand Genomes Project) was downloaded from dbVar (http://www.ncbi.nlm.nih.gov/news/11-04-2014-1000-genomes-phase-3-data-dbvar/) and converted to BEDPE with 25bp of bi-directional slop added. Both the truth and caller-derived SV sets were further filtered with the low complexity region file (Li, 2014), the high coverage region file provided by Lumpy (Layer et al., 2014), and the NA12878 NIST indel calls (http://www.nist.gov/mml/bbd/ppgenomeinabottle2.cfm). These sequential filters were done with the ‘pairToBed -type neither’ command in bedtools. The benchmarks for NA12878 followed the same procedures as the simulations with the exception of mobile element insertions (MEIs). We used ‘pairToPair –type either –slop 50 –is’ for true positives and ‘pairToPair –type neither –slop 50 –is’ for false positives because the MEIs in the truth set are represented with single intervals.

The CHM1 publicly available structural variant call set (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation) was downloaded with the accompanying 101bp Illumina reads (SRX533609). We converted the CHM1 bed files to BEDPE files adding 50bp of bi-directional slop. Without slop, the concordance of the tools tested here and the CHM1 datasets was too low. The Illumina reads were aligned to the b37 (GATK resource bundle) reference genome using BWA mem version 0.7.10-r868-dirty (Li, 2013). The aligned reads were sorted and duplicates were removed using Samtools version 0.1.19-44428cd (Li et al., 2009). WHAM, Lumpy, and Delly were run in default mode over the Illumina BAM files. The benchmarks used the same methods as previously described without the filtering steps as many of the CHM1 calls are in low complexity regions.

6.6.5 Biological datasets

All biological datasets are publically available on Sequence Reads Archive (SRA). The eleven resequenced pigeons used in the association study can be found under SRA054391. The five recessive red birds have the following SRA ids: SRS346872, SRS346882, SRS346899, SRS346902, SRS346883 and the nonrecessive red birds (backgrounds) have the following SRA ids: SRS346895, SRS346873, SRS346870, SRS346896, SRS346874, SRS346877. Each re-sequenced bird has a depth of coverage ~10x from paired-end Illumina reads. The reads were aligned to the pigeon assembly (GenBank assembly accession: GCA_000337935.1). The viral dataset, including the parental (SRS812401) and adapted strain (SRS812403), are associated with SRP051821.

WHAM was run over both the viral and pigeon datasets in default mode. In the pigeon genome-wide association study, we removed sites where there were no-calls (missing genotypes). Sites with LRT values less than 1.5 were also excluded for the Manhattan plot for the purpose of visualizing the dataset. We also removed sites with less than three reads supporting the start position. In the VACV dataset, we removed inverted terminal repeats (10kb on either end of the genome) to avoid false positives from reads mapping to complementary regions of the genome. We then excluded sites where the start position was supported by fewer than 50 reads. Similarly, we discarded Lumpy calls that had fewer than 50 reads supporting a SV. Increasing the filter to 100 removed almost all of the spurious calls from Lumpy.

The genotype-phenotype association test for the *e1* allele using Delly data was done using GPAT++ (Kronenberg). WHAM's likelihood ratio test is implemented in pFst using the "count" setting. Delly was run over the pigeon BAM files using all four modes (DEL, DUP, INV, TRA). These different call sets were merged (union) and passed to pFst (Kronenberg).

6.6.6 PCR validation in the poxvirus dataset

The ΔE3L vaccinia virus was passaged 10 times in primary human fibroblasts at a multiplicity of infection of 0.1 for 48 hours (see (Elde et al., 2012) for details). Deep sequencing of viral populations was performed on libraries prepared from genomic viral DNA isolated from either the parental ΔE3L or an adapted strain after 10 serial passages using the Nextera XT DNA sample prep kit (Illumina, Inc., San Diego, CA, USA). Barcoded libraries were combined and sequenced on a single lane using an Illumina MiSeq instrument. Reads were mapped to the vaccinia virus Copenhagen strain reference genome (accession M35027.1; modified on poxvirus.org) using BWA mem in default mode, duplications were removed using samtools, and SVs in the adapted and parental strains were called with both WHAM and Lumpy.

PCR primers were designed to amplify products across the potential breakpoints identified by WHAM. The K3L breakpoint was amplified using primers K3L break F (5' GGGATAAACTGGTAGGGAAAACTGTAAAAG 3') and K3L break R (5' CAGAGTGAGGATAGTCAAAAAGATAAATGTATAG 3'). The E3L deletion junctions were amplified using E2L int F (5' GGAGCTACAGTTCTTGGC 3'), E4L int R (5' CCTTCGCTATCTCTTATTCGG 3'), and 46731R (5' CTAGCGTACGATCGCTTCTAG 3'). The resulting products were Sanger sequenced and aligned to the reference genome using blastn (NCBI).

6.6.7 Software

WHAM and all associated software can be found on github (https://github.com/jewmanchue/WHAM), documentation is on the wiki (http://jewmanchue.github.io/WHAM/). For community support, please post questions to Biostars.org (Parnell et al., 2011).

6.6.8 Acknowledgement

We would like to acknowledge Gabor Marth, Aaron Quinlan, Ryan Layer, Ryan Abo, and Erik Garrison for their helpful suggestions and critique. Brad Chapman provided extensive feedback for WHAM's classifier and benchmarked WHAM on the Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge. We would also like to thank Carson Holt for rapidly processing the Illumina data.

6.6.9 Online funding statement

We gratefully acknowledge the support of the NIH (R01GM104390 to M.Y.; GM090042 to N.C.E.; T32AI055434 to K.R.C.; T32GM07464 to Z.N.K and E.J.O; 1TL1TR001066 to E.J.O) and the NSF (DEB-1149160 to M.S). N.C.E. is a Pew Scholar in the Biomedical Sciences and Mario R. Capecchi Endowed Chair in Genetics.

6.7 References

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R. a, Hurles, M.E., and McVean, G. a (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M. a, Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G. a (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Andersson, D.I., and Hughes, D. (2009). Gene Amplification and Adaptive Evolution in Bacteria. Annu. Rev. Genet. *43*, 167–195.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, A., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Beattie, E., Denzler, K.L., Tartaglia, J., Perkus, M.E., Paoletti, E., and Jacobs, B.L. (1995). Reversal of the interferon-sensitive phenotype of a vaccinia virus lacking E3L by expression of the reovirus S4 gene. J. Virol. *69*, 499–505.

Bickhart, D.M., Hutchison, J.L., Xu, L., Schnabel, R.D., Taylor, J.F., Reecy, J.M., Schroeder, S., Van Tassell, C.P., Sonstegard, T.S., and Liu, G.E. (2015). RAPTR-SV: a hybrid method for the detection of structural variants. Bioinformatics *31*, 2084–2090.

Brennan, G., Kitzman, J.O., Rothenburg, S., Shendure, J., and Geballe, A.P. (2014). Adaptive gene amplification as an intermediate step in the expansion of virus host range. PLoS Pathog. *10*, e1004002.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2014). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science *327*, 302–305.

Chapman, B. (2015). Blue Collar Bioinformatics. Retrieved July 2, 2015, from http://bcb.io/

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., and Weinstock, G. (2014). TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. Genome Res. *24*, 310–317.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Domyan, E.T., Guernsey, M.W., Kronenberg, Z., Krishnan, S., Boissy, R.E., Vickrey, A.I., Rodgers, C., Cassidy, P., Leachman, S. a, Fondon, J.W., et al. (2014). Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr. Biol. *24*, 459–464.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics *9*, 11.

Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., and Malik, H.S. (2012). Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell *150*, 831–841.

Elliott, K.T., Cuff, L.E., and Neidle, E.L. (2013). Copy number change: evolving views on gene amplification. Future Microbiol. *8*, 887–899.

Erlandson, K.J., Cotter, C.A., Charity, J.C., Martens, C., Fischer, E.R., Ricklefs, S.M., Porcella, S.F., and Moss, B. (2014). Duplication of the A17L Locus of vaccinia virus provides an alternate route to rifampin resistance. J. Virol. *88*, 11576–11585.

Faust, G. (2015). GregoryFaust/SVsim. Retrieved July 2, 2015, from https://github.com/GregoryFaust/SVsim

Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503–2505.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. *44*, 445–477.

Hart, S.N., Sarangi, V., Moore, R., Baheti, S., Bhavsar, J.D., Couch, F.J., and Kocher, J.-P. a (2013). SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. PLoS One *8*, e83356.

Homer, N. (2015). Nh13/DWGSIM. Retrieved July 2, 2015, from https://github.com/nh13/dwgsimIllumina

Kemena, C., and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics *25*, 2455–2465.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. Genet. Epidemiol. *34*, 479–491.

Kronenberg, Z. (2015). Jewmanchue/vcflib. Retrieved July 2, 2015, from https://github.com/jewmanchue/vcflib/wiki

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. Nucleic Acids Res. *41*, D936–D941.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv1303.3997 *00*, 1–3.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843–2851.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and others (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., Cao, H., Wu, B., Huang, S., Shao, H., et al. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat. Biotechnol. *29*, 725–732.

Marschall, T., Costa, I.G., Canzar, S., Bauer, M., Klau, G.W., Schliep, A., and Schönhuth, A. (2012). CLEVER: clique-enumerating variant finder. Bioinformatics *28*, 2875–2882.

Marschall, T., Hajirasouliha, I., and Schönhuth, A. (2013). MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. Bioinformatics *29*, 3143–3150.

McCarroll, S. a, and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. Nat. Genet. *39*, S37–S42.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Michaelson, J., and Sebat, J. (2012). forestSV: structural variant discovery through statisical learning. Nat. Methods *9*, 819–821.

Mills, R.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. *16*, 1182–1190.

Mimori, T., Nariai, N., Kojima, K., Takahashi, M., Ono, A., Sato, Y., Yamaguchi-Kabata, Y., and Nagasaki, M. (2013). iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. BMC Syst. Biol. *7 Suppl 6*, S8.

Narzisi, G., O’Rawe, J. a, Iossifov, I., Fang, H., Lee, Y., Wang, Z., Wu, Y., Lyon, G.J., Wigler, M., and Schatz, M.C. (2014). Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat. Methods *11*, 1–7.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Onishi-Seebacher, M., and Korbel, J.O. (2011). Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. Bioessays *33*, 840–850.

Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M. a, Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol. *11*, R52.

Parnell, L.D., Lindenbaum, P., Shameer, K., Dall'Olio, G.M., Swan, D.C., Jensen, L.J., Cockell, S.J., Pedersen, B.S., Mangan, M.E., Miller, C. a, et al. (2011). BioStar: an online question & answer resource for the bioinformatics community. PLoS Comput. Biol. *7*, e1002216.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Perry, G., Yang, F., and Marques-Bonet, T. (2008). Copy number variation and evolution in humans and chimpanzees. Genome Res. *18*, 1698–1710.

Platinum Genomes. (2015). Retrieved July 2, 2015, from http://www.illumina.com/platinumgenomes/

Quinlan, A., and Clark, R. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. *20*, 623–635.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Romero, D., and Palacios, R. (1997). Gene amplification and genomic plasticity in prokaryotes. Annu. Rev. Genet. *31*, 91–111.

Sandegren, L., and Andersson, D.I. (2009). Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. *7*, 578–588.

Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., et al. (2013). Genomic diversity and evolution of the head crest in the rock pigeon. Science *339*, 1063–1067.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.

Sindi, S.S., Onal, S., Peng, L.C., Wu, H.-T., and Raphael, B.J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol. *13*, R22.

Slabaugh, M.B., Roseman, N.A., and Mathews, C.K. (1989). Amplification of the ribonucleotide reductase small subunit gene: analysis of novel joints and the mechanism of gene duplication in vaccinia virus. Nucleic Acids Res. *17*, 7073–7088.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.

Steinberg, K.M., Schneider, V.A., Graves-lindsay, T.A., Fulton, R.S., Agarwala, R., Huddleston, J., Shiryev, S.A., Morgulis, A., Surti, U., Warren, W.C., et al. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. *20*, 2066–2076.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 1–2.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138.

Wong, K., Keane, T.M., Stalker, J., and Adams, D.J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. *11*, R128.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.

Table 6.1. The number of NA12878 calls before and after filtering. The rows show how many calls are removed if each filter is applied independently. The “all filters combined” row was the data used for the benchmarks presented in Figure 6.1.

	Delly	LUMPY	WHAM	Phase III NA12878
Total calls	53,946	11,907	579,004	2,597
LCR filter	43,388	6,341	274,787	2,149
High coverage filter	52,427	8,101	579,004	2,520
INDEL filter	50,666	11,004	510,994	2,574
All filters combined	40,652	3,678	250,730	2,079
Deletions after filters	40,652	2,166	6,935	658

Table 6.2. WHAM and Lumpy breakpoint positions in the adapted vaccinia virus population (related to Fig. 6.5). WHAM and Lumpy breakpoint positions in the adapted vaccinia virus genome, corresponding to the breakpoints shown in Fig. 6.4. WHAM split-read (SR) and mate-pair (MP) read support is listed for each position. Asterisks (*) indicate breakpoints for which the breakpoint position is the same as the Sanger sequencing verified breakpoint.

WHAM breakpoint	WHAM breakpoint position	WHAM SR count	WHAM MP count	Lumpy breakpoint position
1	31725*	977	73	31725*
2	30296*	456	100	30295
3	50913*	0	6	N/A
4	46840	1	160	46746
5	46731*	344	65	N/A
6	51483*	314	62	51482

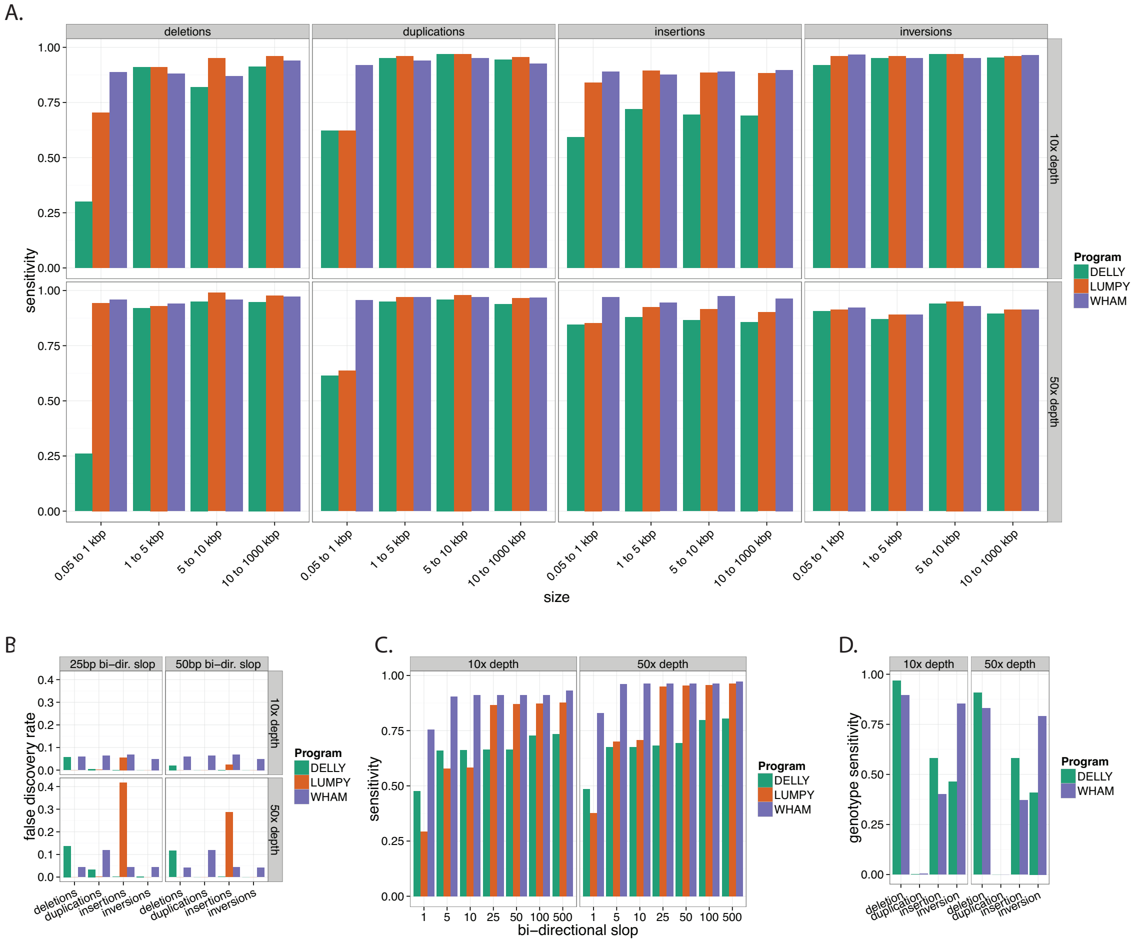
Table 6.3. The factors used to classify the SV type. All factors reported in the VCF "AT" info field are the fraction of the reads in the pileup falling into one or more of the 14 categories described above. During the training step, the importance of each factor is reported as shown above. The importance is the relative weight each attribute contributes to the classification.

Category	Notes	Importance
0. Both mates mapped	NOT USED	NOT USED
1. Discordant	Not paired	0.056
2. Mate not mapped	SAM bit flag 8	0.0
3. Mates mapped to same strand	Mate pairs on same strand	0.089
4. Mates on different seqids	RefID != MateRefID	0.148
5. Number of split reads	Contains the 'SA' optional tag	0.078
6. Split read (fragment 1) on same strand as mate	combination of SAM flag and SA tag	0.071
7. Split read (fragment 2) on same strand as mate	combination of SAM flag and SA tag	0.071

Table 6.3 continued

Category	Notes	Importance
8. Split read (fragment 1) and read two (fragment 2) on same strand	combination of SAM flag and SA tag	0.13
9. Internal insertion	Cigar string contains ‘I’ operation	0.0002
10. Internal deletion	Cigar string contains ‘D’ operation	0.0135
11. Mates mapped too close	Insert < (2.5 * sd average insert size)	0.024
12. Mates mapped to far	Insert > (2.5 * sd average insert size)	0.101
13. Everted pairs	Position and orientation of mate pairs	0.075
14. Relative depth	Depth at position for each sample relative to their mean depth	0.137

Figure 6.1. Sensitivity and false discovery rates (FDR) for simulated data.
The specificity and FDR of Delly, Lumpy, and WHAM for simulated deletions, duplications, insertions, and inversions. The sensitivity is measured for each category at depths of 10x and 50x. SVs ranging from 50bp to 1Mb are grouped into four left-closed size intervals. A) The sensitivity of the three tools is faceted on size, depth, and SV type. At 10x WHAM has noticeably better sensitivity for deletions and duplications in the smallest size class. WHAM's sensitivity is higher than Delly and Lumpy for insertions at 10x and gains sensitivity at 50x. B) The FDR for each type of SV faceted by depth and the amount of slop added to each confidence interval. In the 25bp slop category, each confidence interval was extended in both directions by 25bp. At 10x depth, WHAM has the highest FDR across all SV classes and Lumpy has the lowest. At 50x, Delly has heightened FDR for deletions and Lumpy has a much higher FDR for insertions. Shrinking the confidence intervals increases the FDR for Delly and Lumpy, but not WHAM. C) Breakpoint sensitivity for deletions. The confidence intervals, provided by the three tools, are ignored and slop is incrementally added to the predicted breakpoints. WHAM has the highest sensitivity when 1-10bp of slop is added. D) Genotype sensitivity for the homozygous nonreference simulated SVs. Delly and WHAM have similar sensitivity for deletions and duplications while both tools fail to correctly genotype duplications.


A.
deletions
duplications
insertions
inversions
10x depth
50x depth
sensitivity
1.00
0.75
0.50
0.25
0.00
0.05 to 1 kbp
1 to 5 kbp
5 to 10 kbp
10 to 1000 kbp
size
Program
DELLY
LUMPY
WHAM
B
25bp bi–dir. slop
50bp bi–dir. slop
false discovery rate
0.4
0.3
0.2
0.1
0.0
deletions
duplications
insertions
inversions
C.
bi–directional slop
1
5
10
25
50
100
500
D.
genotype sensitivity
deletion
duplication
insertion
inversion

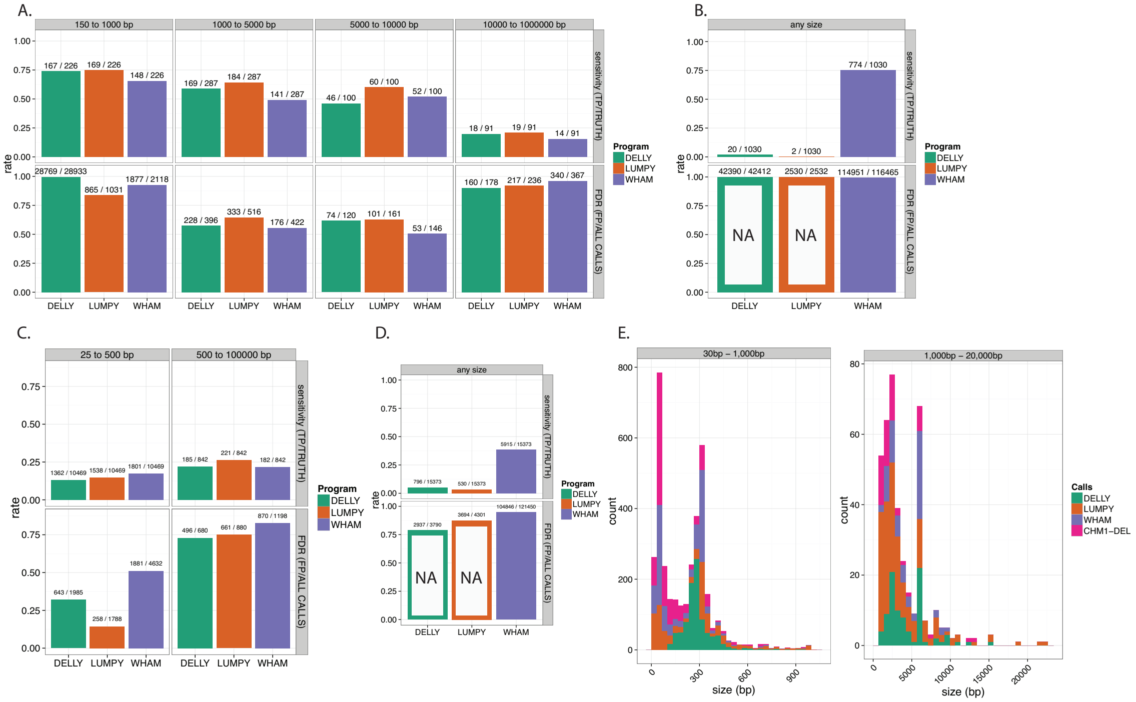


Fig. 6.2. Benchmarking Delly, Lumpy, and Wham against NA12878 and CHM1 datasets. A) The sensitivity and FDR for filtered NA12878 Phase III deletion calls across four size intervals. The number of true positives and the number NA12878 calls are listed above sensitivity, while the total number of false positives and total calls for each tool are listed above FDR. Most true positives and false positives are within the 150bp to 1,000bp interval. B) The sensitivity and FDR for NA12878 Phase III mobile element insertion (MEI) calls. Unlike the other benchmarks, if either breakpoint of a SV call falls within the MEI truth interval, the call is considered a true positive. Delly and Lumpy calls, regardless of SV type, are intersected with the MEI calls as are WHAM insertion calls. The FDRs for Delly and Lumpy are not meaningful as the call sets were not restricted to one SV type. WHAM insertion calls overlap 75% of the MEI calls in Phase III One Thousand Genomes Project. C) The sensitivity and FDR for all CHM1 deletions. In the 25bp to 500bp interval size, WHAM discovers ~200 more SVs than the other tools, while Lumpy finds more of the SVs in the 500bp to 100Kbp interval. D) The size distribution of the true positive calls that overlap the CHM1 deletions. One thousand true positives were randomly sampled from each tool and the truth set (CHM1-DEL).

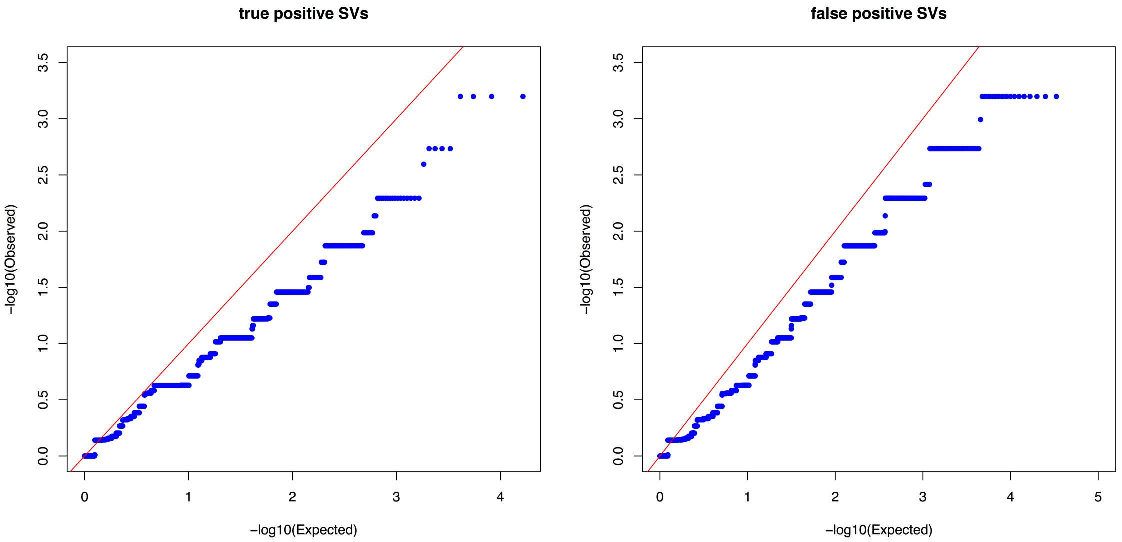


Fig. 6.3. WHAM false positives and true positives share similar p-value distributions. Quantile-quantile plots for WHAM's LRT statistic after conversion to p-values (y-axis). Left panel: The p-values for the structural variants that intersect with the phase III One Thousand Genomes Project dataset (within +/-25bp). Right panel: The p-values for structural variants that do not intersect with the phase III One Thousand Genomes Project dataset. Both the true and false positive SV calls have very similar distributions.

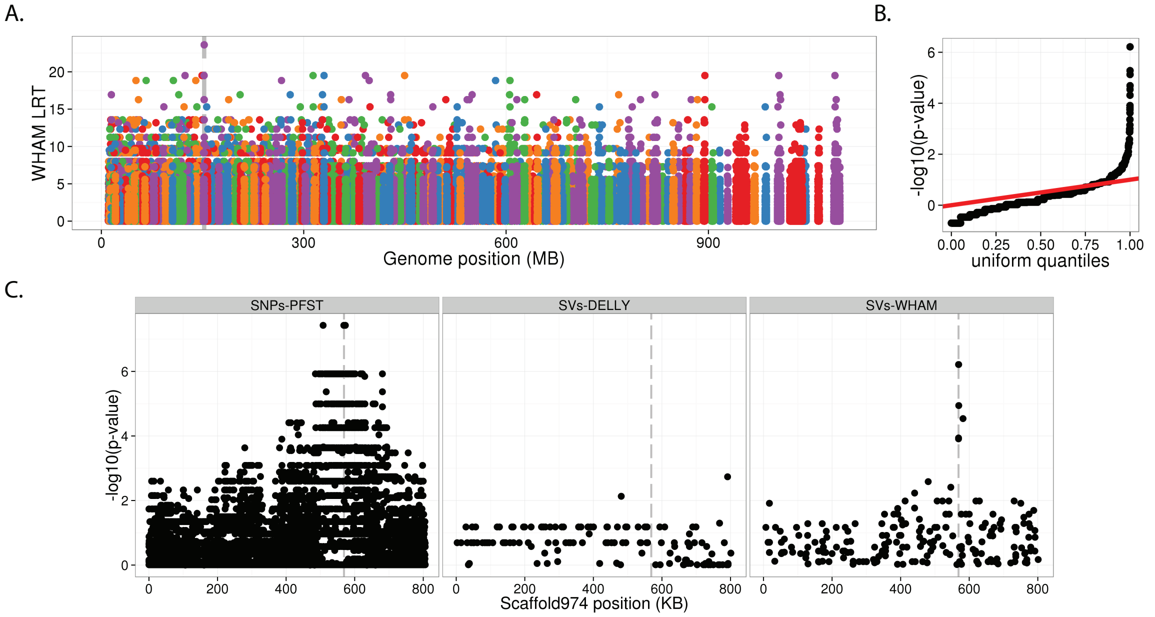
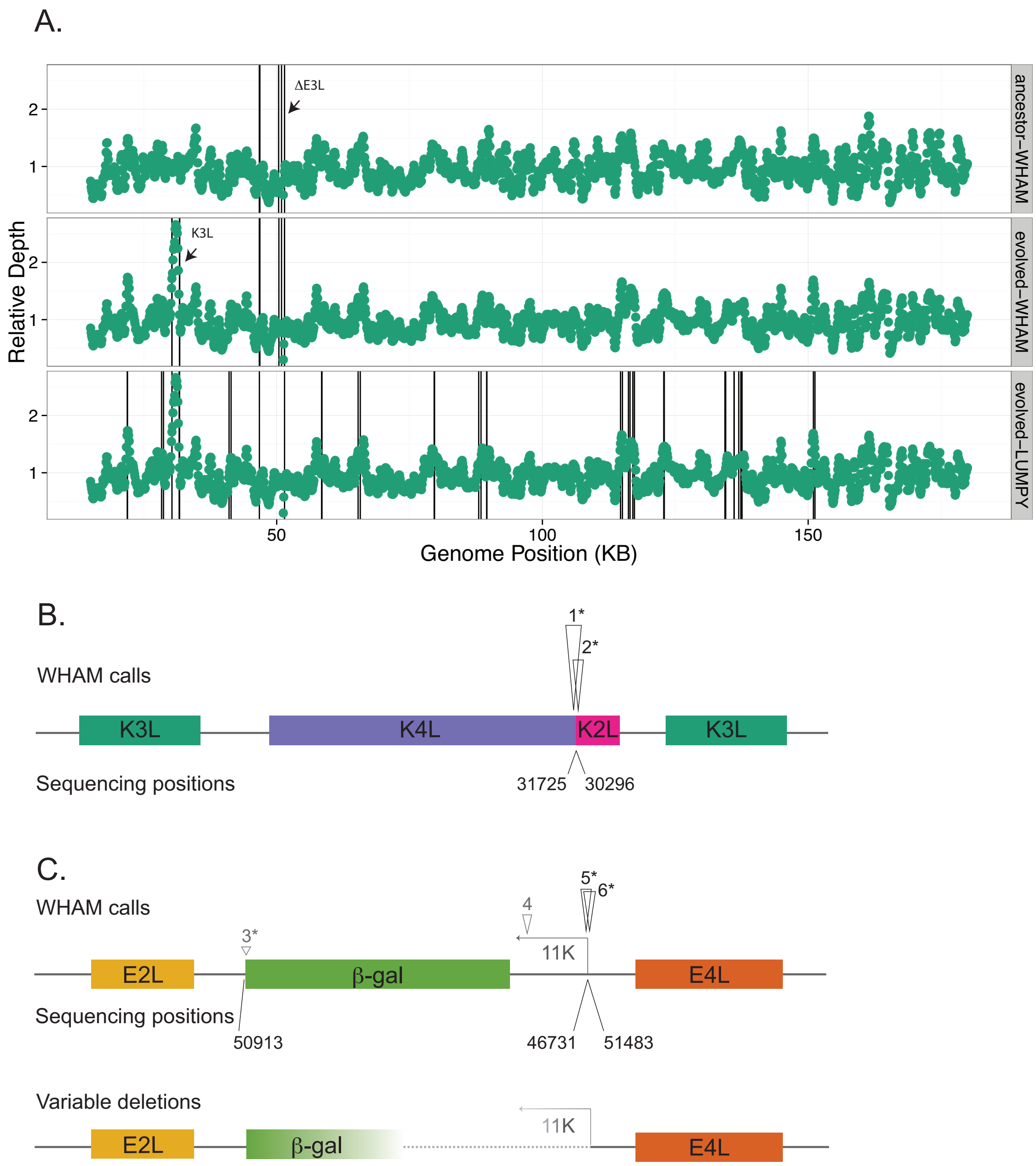


Fig. 6.4. Identification of the *e1* allele using WHAM's LRT. A) WHAM's LRT interrogates allele frequency differences between recessive red and wild-type birds. The colors denote different scaffolds in the pigeon assembly. Scaffolds are sorted by their size in increasing order. The highest LRT score, denoted with the dashed vertical line, is a 10Kb deletion upstream of the *Sox10* gene. The Sox10 transcription factor is important in the melanin synthesis pathway. B) The quantile-quantile plot after converting WHAM's likelihood ratio values to p-values. Only LRT values above 1.5 are shown in A. C) Scaffold974 association tests from SNPs, WHAM SV calls, and Delly SV calls.

Fig. 6.5. WHAM detects structural variation in vaccinia virus populations A) Read depth normalized within each sample is plotted across the ~200kb vaccinia genome (excluding inverted terminal repeats) for either the parental strain (top panel) or an adapted strain (middle and bottom panels, called by WHAM or Lumpy, respectively). Arrows highlight the positions of K3L CNV and E3L deletion. The black lines represent the breakpoints of every SV call after filtering (see Materials and Methods). B) WHAM calls in the adapted strain near the K3L duplication breakpoint are shown as black triangles above the viral genes in colored boxes. The height of the triangle represents split-read (SR) count supporting the call. Sanger sequencing positions relative to the reference sequence are listed below. Asterisks (*) indicate WHAM calls that match the exact breakpoint determined by Sanger sequencing (see Table S1 for WHAM and Lumpy breakpoints). C) WHAM calls in the adapted strain near the E3L deletion are shown above the genes, and Sanger sequence confirmed positions below, as in B. The arrow indicates the position of the 11K promoter driving β-gal expression. For breakpoints in grey, the height of the triangle indicates the relative mate-pair count from WHAM, as these positions do not have SR support.


A.
ΔE3L
K3L
Relative Depth
2
1
ancestor–WHAM
evolved–WHAM
evolved–LUMPY
50
100
150
Genome Position (KB)
B.
WHAM calls
1*
2*
K3L
K4L
K2L
K3L
Sequencing positions
31725
30296
C.
WHAM calls
3*
4
5*
6*
11K
E2L
β-gal
E4L
Sequencing positions
50913
46731
51483
Variable deletions
E2L
β-gal
11K
E4L

CHAPTER 7

CONCLUSION

The work presented in this thesis is an attestation that High Throughput Sequencing (HTS), even with its sizable error rates, is suitable for Genotype-Phenotype Association (GPA) studies. In this dissertation, I have presented examples of successful HTS GPA studies in nonmodel organisms. These studies were greatly facilitated by the novel GPA methods I have developed, GPAT++ and WHAM. These two tools, when applied to both structural-variant and SNP data, can find associations between genetic variants and phenotypes. In this chapter, I will discuss how these tools have enabled new lines of biological inquiry, how new types of HTS data will improve GPA methodology, and how new types of data will affect GPA studies.

7.1 Enabling biological inquiry with GPAT++ and WHAM

Both WHAM and GPAT++ were built over the course of several biological projects. Each tool was originally designed for a single purpose, but they evolved into fully developed and general software packages that can be used for many kinds of GPA studies. In addition to being broadly applicable, the interfaces of these tools are user friendly. Often, collaborators who use our tools have limited experience with command line environments. Therefore, it is critical to provide sufficient user documentation. Both GPAT++ and WHAM have publically available wikis, documenting usage statements and providing use case scenarios. Additionally, we have provided downstream tools that help biologists interpret GPAT++ and WHAM output.

Both of our GPA tools are being used locally and globally. Within the research community at the University of Utah, our methods have been applied to pigeon, virus, whale, pine fungus, *C. elegans*, human exome data, a human chromosome transferred into a mouse cell line, and pathogenic strains of *Streptococcus phenomena*. Outside of collaborations, 153 people have accessed GPAT++ and 54 people have accessed WHAM in April. The global distribution of users is dispersed across many continents. We hope to continue to provide support and novel GPA methods in both WHAM and GPAT++ as our user base grows.

7.2 Third-generation sequencing technologies will improve GPA

Third-generation sequencing technologies have the potential to drastically improve the efficacy of GPA. These technologies, unlike second-generation sequencing, require fewer enzymatic steps before sequencing. For example, PacBio and Oxford nanopore do not require PCR amplification of genomic DNA. Third-generation sequencing platforms are also advantageous because they produce read lengths an order of magnitude longer than second-generation sequencing. Even with high error rates, these long reads can be mapped back to the reference genome with higher accuracy. Improved read mapping reduces the number of false variants that are included in GPA studies

Third-generation sequencing technologies will allow GPA studies to venture into complex and repetitive regions of the genome. Expansions of simple repeats have been associated with over thirty different human diseases (Mirkin 2007). Many of these expansions are longer than second-generation reads, but not third-generation reads. Similarly, the genetic variation contained within LINE-1 and ALU elements is difficult to characterize with short-read technologies. Longer reads will allow both clinicians and biologists to study how complex regions of the genome cause disease and phenotype variation.

Structural variation (SV) GPA studies will greatly benefit from third-generation sequencing. Structural variants are difficult to detect when they are not entirely contained within a read; take for example novel insertions. If the inserted sequence is longer than a read, it cannot be characterized with mapping-based methods. Currently PacBio reads are, on average, long enough to capture the entire sequence of an ALU insertion, the most common DNA sequence in the human genome. Localizing breakpoints of SVs will become less burdensome with third-generation sequencing. SVs that are partially contained within a read lower that read's mapping score or cause the read to not map entirely. Third-generation sequencing reads carrying partial SV sequences are more likely to map correctly, which allows for the breakpoint of the SV to be found. Third-generation sequencing will provide a way to properly identify SVs and characterize SV diversity within the human population.

While third-generation sequencing technologies will greatly improve GPA, it has not been widely adopted due to fiscal reasons. These technologies are too expensive for population-level sequencing projects, required for GPA. However, hybrid approaches might be a solution. Individual genomes could be sequenced at high depth with second-generation sequencing and at very low depths with third-generation sequencing technologies. These data could drastically improve GPA for both structural and single nucleotides variants while being finically feasible.

7.3 Future computational challenges for GPA

The ever-growing size of HTS population-level GPA data requires novel computational methods. The N+1 problem is an example of how the number of HTS samples is creating a computational nightmare. N+1 is a concept in which a single new genome is sequence, but it needs to be analyzed with a cohort of other genomes. As more and more genomes are sequenced, the computational cost grows rapidly. Representing the known genetic variation in a format that allows for new genomes to be analyzed, in context of current genomes, without processing past genomes is paramount. On this front, several groups are working on graph-based representation of genetic variation. HTS reads from new genomes can be aligned to a graph containing both the reference genome and the previously characterized genetic variation. This process can be repeated many times, saturating the genetic variation represented in a single graph.

If graph-based genetic variant representation is proven to be effective, GPA methods will need to be updated to conduct association testing from the graph data structure. This is one area where GPAT++ could be expanded in the future. GPAT++ would need to be able to traverse the graph and identify paths enriched in the target cohort less frequently in the background.

Another computational task warranting further efforts is GPA using population-level whole genome *de novo* assemblies. By creating *de novo* assemblies, the nature and sequence of many small structural variants can be resolved. The challenge will then be to compare these assemblies to one another or to a reference genome. New GPA methodology will need to handle the comparison of thousands of assembled sequences across many individuals.

All of the computational challenges discussed provide exciting opportunities for the next wave of Ph.D. students in the field of Computational Biology. There is much we still do not know about how genetic variation causes phenotypic variation and hopefully continual advances in HTS will allow us to find GPA between complex phenotypes and novel genetic variants.

7.4 References

Mirkin, Sergei M. 2007. Expandable DNA repeats and human disease. Nature. *447(7147)*, 932–40.