A novel sequencing method to explore two molecular mechanisms of rapid adaptation

Keisha Dawn Carlson

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Christine Queitsch, Chair

Maitreya Dunham

Jay Shendure

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

A novel sequencing method to explore two molecular mechanisms of rapid adaptation

Keisha Dawn Carlson

Chair of the Supervisory Committee
Associate Professor Dr. Christine Queitsch
Department of Genome Sciences

To survive in an ever-changing world, organisms need to rapidly adapt to new environments. In my dissertation, I address two molecular mechanisms that may contribute to rapid adaptation to new environments and help to address the "missing heritability" of complex traits. The first mechanism is short tandem repeats (STRs), short sequence units of two to ten nucleotides repeated head to tail, which mutate 10 to 10,000x faster than non-repetitive regions and have been shown to increase mutation rate under environmental stress. These highly variable, phenotypically important genetic elements have remained inaccessible to high-throughput analysis with short read sequencing and therefore have been excluded from genome-wide analyses of genotype-phenotype associations. I developed a novel method MIPSTR to accurately genotype STRs genome-wide across many individuals. The second mechanism is buffers of genetic variation, *i.e.* molecules or pathways that influence the penetrance of standing genetic variation by buffering its effect on phenotype. This mechanism could also be thought of as an epistatic interaction with loci across the genome. If a buffering mechanism is challenged or breaks, previously cryptic genetic variation will be revealed. In order to consider the robustness (*i.e.* buffering capacity) of an organism when associating genetic variants to complex traits, we need to measure robustness of individuals. MIPSTR can assess somatic STR variation within individuals as a measure of genome stability and thereby robustness. This new approach has allowed me to address two molecular mechanisms potentially important for adaptation and "missing heritability" in novel ways.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I wish to offer my thanks and gratitude to:

My advisor, Christine Queitsch, who always managed to keep me enthusiastic about science even through frustrating times and who I respect as a brilliant scientist and caring friend.

Members of the Queitsch lab, past and present, who offered advice, support, and insight: Jen Lachowiec (who in particular helped me to be a productive and thoughtful scientist through ceaseless encouragement and by setting an excellent example), Karla Schultz, Sanna Sullivan, Max Press, Michael Dorrity, Cris Alexandre, Tzitziki Lemus, Pauline Rival, Kerry Bubb, Alex Mason, Beth Morton, James Urton, Janne Lempe, Soledad Undurraga, and Jacob Bale.

My committee: Maitreya Dunham, Jay Shendure, Evan Eichler, and Ben Kerr for thoughtful feedback, vital collaborations, and consistent encouragement.

Genome Sciences graduate students and postdocs for input, advice, commiseration, and laughs.

My rugby teammates for giving me an outlet for aggression stemming from failed experiments.

My friends for support, love, and many good times.

Phil for love and for helping me mellow out and go with the flow.

My family for constant support and unconditional love.

CHAPTER ONE
INTRODUCTION

To survive in an ever-changing world, organisms need to rapidly adapt to new environments. Sessile organisms like plants, which cannot migrate to more suitable environments, have a particular need to adapt rapidly. Plants also rely on environmental cues for healthy growth and development and misinterpretation of such cues can be devastating (Fehér 2014). In order for organisms to adapt they must show a diversity of phenotypes on which natural selection can act. Underlying this natural variation in phenotype is variation in DNA sequence, as well as the environment and stochastic noise. In order to understand how organisms, and in particular plants, can create this phenotypic diversity and adapt rapidly to novel environments, we study the model plant *Arabidopsis thaliana*.

A. thaliana grows throughout the northern hemisphere and the global population has a similar level of genetic diversity to human populations (Cao et al. 2011). Different strains can inhabit diverse environments and have adapted to do so in a variety of traits such as timing of the transition from vegetative growth to reproductive growth (flowering time) and circadian rhythm (Lewandowska-Sabat et al. 2012; Dittmar et al. 2014; Rival et al. 2014). In fact, some strains are so uniquely adapted to their environment that natural alleles of key developmental genes from other strains are incompatible in their genetic background (Undurraga et al. 2012). In addition to vast natural variation, *A. thaliana* has numerous genomic, genetic, and molecular tools available. There is a high-quality genome for *A. thaliana*, hundreds of re-sequenced genomes of diverse strains, a large insertion mutant library, RNAi knockdown techniques, and easy transformation protocols. With all of these tools in hand, *A. thaliana* is an ideal model for studying molecular mechanisms underlying rapid adaptation to new environments. Exploring such mechanisms in *A. thaliana* can not only help us to better understand evolutionary phenomena such as speciation, but can also translate to increasing crop yield in harsh or sub-optimal growth environments.

A classic example of rapid adaptation to novel environments is that of the peppered moths during the industrial revolution (Cook and Saccheri 2013). The predominantly salt-and-pepper color of these moths before the industrial revolution allowed them to blend in with the lichen on the trunks of trees, making them difficult for predators to spot. As a result of the rise of industrial power plants, many of the trees lost lichen from their trunks and turned the color of

soot. Very soon after, moths with soot colored wings rose to high frequency because in this new environment, the light moths were easy prey for birds and the salt-and-pepper trait was selected against (Cook and Saccheri 2013).

In this case, the molecular mechanism underlying the rapid adaptation is known; the peppered phenotype is a recessive trait and a dominant mutation causes the soot colored wings via an increase in melanin production (Cook and Saccheri 2013). This mechanism, a dominant mutation, can quickly lead to an increase in frequency of the corresponding trait within a population under a strong selective pressure (Cook and Saccheri 2013). However, most mutations are not dominant and in fact have little or no effect on phenotype at all (Gruber et al. 2012; Rutter et al. 2012). Additionally, many complex traits of interest are not controlled by a single genetic locus. Often, multiple loci interact epistatically, with each other and with the environment, in addition to experiencing stochastic noise before translating into phenotype of an organism (Bloom et al. 2013; Eichler et al. 2010; Manolio et al. 2009; Queitsch et al. 2012b). This leads us to the conclusion that there must be other molecular mechanisms underlying rapid adaptation to novel environments.

To understand these mechanisms, we first need to understand how genotypes translate to phenotypes and how these genotypes and phenotypes are inherited. The peppered moth story turns out to be very straight forward, but as our field is quickly learning, complex traits are rarely straightforward. In fact, the field is left with “missing heritability”, or traits known to be partially heritable based on family studies for which the underlying genetic cause cannot be identified. We suggest this “missing heritability” is at least partially due to 1) the exclusion of some classes of genetic variants, which are inaccessible with current short read sequencing technologies, in current association and mapping studies and 2) the lack of methods for analyzing the effects of epistasis by considering combinations of genetic loci and environmental conditions.

In my dissertation, I address two molecular mechanisms, one from each of the two categories above, which may contribute to rapid adaptation to new environments and help to address the “missing heritability” of complex traits. The first mechanism is short tandem repeats (STRs), short sequence units of two to ten nucleotides repeated head to tail, which mutate 10 to 10,000x faster than non-repetitive regions and have been shown to increase mutation rate under environmental stress (Gemayel et al. 2012b; Rando and Verstrepen 2007). STRs, also known as microsatellites, exist in coding and promoter regions of genes where they are likely functional.

We hypothesize the high, environmentally responsive mutation rate of STRs allows for increased phenotypic diversity on which natural selection can act. These highly variable, phenotypically important genetic elements have remained inaccessible to high-throughput analysis with short read sequencing (**Fig. 1.1**). This inaccessibility results from high technical error rates during amplification and low sequence complexity leading to inaccurate or uninformative mapping (Press et al. 2014). Therefore, STRs have been excluded from genome-wide analyses of genotype-phenotype associations. In order to address the role of STRs in the genotype-phenotype map, we need a new method for accurate, high-throughput genotyping of STRs across many individuals.

Figure 1.1 Sequences from 1001 genomes project all show reference bias. The gene HSFa1a has a highly variable triplet repeat in its coding sequence highlighted here. The Col-0 reference sequence is shown in the top row. Each of the other re-sequenced strains shown here were reported to have the same STR unit number genotype as the reference, probably due to reference bias. The true STR unit number genotypes from Sanger sequencing data is overlaid.

The second mechanism is buffers of genetic variation, *i.e.* molecules or pathways that influence the penetrance of standing genetic variation by buffering its effect on phenotype. This mechanism could also be thought of as an epistatic interaction with loci across the genome. The primary example of such a buffering mechanism is the protein chaperone and evolutionary capacitor, HSP90, which can help clients fold and function despite genetic variation (Queitsch et al. 2002; Sangster et al. 2008a). Such buffering mechanisms can help to maintain genetic variation in a population by buffering its effect on phenotype, rendering it effectively neutral (Queitsch et al. 2002; Sangster et al. 2008a). For example, proteins that are clients of HSP90

experience relaxed purifying selection and therefore acquire more non-synonymous mutations than non-clients over time (Lachowiec et al. 2013, 2014). HSP90 can help its clients to fold and function properly even in the presence of slightly deleterious mutations. If a non-client acquired a similar mutation, this protein could no longer fold or function correctly, so the variant would be purified out of the population. However, if a buffering mechanism is challenged or breaks, this previously cryptic genetic variation will be revealed. In the case of HSP90, if an organism undergoes environmental stress, many non-client proteins begin misfolding and titrating HSP90 away from its normal client base. Then variants in those clients that were previously cryptic could lead to novel phenotypes and if those phenotypes are advantageous, to rapid adaptation (Queitsch et al. 2002; Sangster et al. 2008a).

Although HSP90 is the best-studied example of a genetic buffering mechanism, there are many other potential buffering mechanisms such as redundancy in genetic networks, chromatin remodeling factors, and small RNAs (Wagner 2000; Bergman and Siegal 2003; Levy and Siegal 2008; Leclerc 2008; Ciliberti et al. 2007; Li et al. 2009; Posadas and Carthew 2014; Ito et al. 2011; Hornstein and Shomron 2006). For example, small RNAs can finely tune expression levels of target genes even when there is genetic variation in the target gene promoter or gene body itself (Li et al. 2009; Tzitziki Lemus, personal communication). We propose that if the extent to which an organism is buffered, or its robustness, can be taken into account as an interaction with genetic variation, more of the "missing heritability" of complex traits will be explained.

In addition to buffering standing genetic variation, we hypothesize that these robustness regulators will also buffer the effects of new mutations. Under our model of the role of "missing heritability" in complex traits, many of the causative variants are young, rare mutations or de novo, private mutations that have a stronger effect in less robust organisms. In my graduate work, I began to address whether HSP90 can buffer new mutations as well as standing genetic variation. I found that new mutations are more penetrant in HSP90 reduced plants than in wild-type controls. Still, this work only applies to HSP90 and more work is needed to understand the effects of naturally varying levels of robustness on mutation penetrance.

In order to consider the robustness (*i.e.* buffering capacity) of an organism when associating genetic variants to complex traits, we need to measure robustness of individuals. Others and we have observed that reduction in HSP90 leads to a decrease in genome stability in multiple organisms from yeast to humans, including increased transposon mobility, mitotic

homologous recombination events, and somatic STR slippage (Mittelman et al. 2010; Specchia et al. 2010a; Chen et al. 2012; Queitsch et al. 2012). We propose that level of robustness is associated with genome stability and to use genome stability as a molecular marker for robustness.

To address STRs and robustness as mechanisms for rapid adaptation to new environments, I developed a sequencing based method called MIPSTR. MIPSTR uses molecular inversion probes (MIPs) to capture STR loci genome wide before sequencing and mapping with a novel strategy. Each probe has a unique DNA tag, allowing us to identify sequences of individual molecules of DNA. This not only allows for very high accuracy (we can get a consensus from many reads per molecule), but also for the identification of somatic STR mutations. The combination of throughput and accuracy of MIPSTR is unprecedented for genotyping STRs. Applying MIPSTR to diverse genomes will allow us to develop our understanding of the role STRs play in phenotype, heritability, and adaptation. Because STRs are so highly mutable, they are an excellent target for identifying somatic variation. MIPSTR can assess somatic STR variation within individuals as a measure of genome stability and thereby robustness. This new approach has allowed me to address two molecular mechanisms potentially important for adaptation and “missing heritability” in novel ways.

CHAPTER TWO

THE OVERDUE PROMISE OF SHORT TANDEM REPEAT VARIATION FOR HERITABILITY[1]

Keywords: short tandem repeats, microsatellites, heritability, epistasis, sequencing technologies

Abstract

Short tandem repeat (STR) variation has been proposed as a major explanatory factor in the heritability of complex traits in humans and model organisms. However, we still struggle to incorporate STR variation into genotype-phenotype maps. Here, we review the promise of STRs in contributing to complex trait heritability, and highlight the challenges that STRs pose due to their repetitive nature. We argue that STR variants are more likely than single nucleotide variants to have epistatic interactions, reiterate the need for targeted assays to accurately genotype STRs, and call for more appropriate statistical methods in detecting STR-phenotype associations. Lastly, somatic STR variation within individuals may serve as a read-out of disease susceptibility, and is thus potentially a valuable covariate for future association studies.

The 'missing heritability' of complex diseases and STR variation.

Complex diseases such as diabetes, various cancers, cardiovascular disease, and neurological disorders cluster in families, and are thus considered to have a genetic component (Manolio et al. 2009; Eichler et al. 2010; Gibson 2011) (Glossary). The identification of these genetic factors has proven challenging; although genome-wide association (GWA) studies have identified many genetic variants that are associated with complex diseases, these generally confer less disease risk than expected from empirical estimates of heritability. This discrepancy, termed the 'missing heritability', has been attributed to many factors (Zuk et al. 2012; Manolio et al. 2009; Eichler et al. 2010; Gibson 2011; Bloom et al. 2013; Heng 2010). A trivial explanation is that shared environments among relatives may artificially inflate estimates of heritability. However, missing heritability may also be due to variants in the human genome that are currently inaccessible at a population scale (Manolio et al. 2009; Eichler et al. 2010). One such class of variation is short tandem repeat (STR) unit number variation. Some have previously

[1] This chapter is accepted in *Trends in Genetics* as "The overdue promise of short tandem repeat variability for heritability" by MO Press, KD Carlson, C Queitsch.

suggested that adding STR variation to existing genetic models would considerably increase the proportion of heritability explained by genetic factors in human disease (Fondon et al. 2008b; Hannan 2010). Three percent of the human genome consists of STRs (Subramanian et al. 2003) and 6% of human coding regions are estimated to contain STR variation (O'Dushlaine et al. 2005; Mularoni et al. 2006). Recently, the first catalog of genome-wide population-scale human STR variation has appeared (Willems et al. 2014), opening up new possibilities for understanding the contribution of STRs to human genetic diseases. This catalog, and similar data sources (Mackay et al. 2012), have appeared only decades after initial calls for the assessment of the role of STRs in phenotypic variation (Kashi et al. 1997), lagging behind surveys of other genomic elements. The initial interest in STRs was due to the discovery of phenomena such as genetic anticipation, which are mediated by the unique features of STRs (Sutherland et al. 1991). It is our hope that new and forthcoming data sources can help us begin to realize the long-deferred promise of STRs for explaining heritability.

STRs consist of short (2-10 bp) DNA sequences (units) that are repeated head-to-tail multiple times. This structure causes frequent errors in recombination and replication that add or subtract units, leading to STR mutation rates that are 10-fold to 10^4-fold higher than those of non-repetitive loci (Eckert and Hile 2009; Legendre et al. 2007). Due to technical barriers, STR variation has until very recently remained inaccessible to genome-wide assessment.

STRs are often conserved (even if their unit number or even sequence changes), especially in coding sequences (Zhao et al. 2014; Schaper et al. 2014; Li et al. 2012; Gemayel et al. 2010). In both humans and the yeast *Saccharomyces cerevisiae*, promoter regions are known to be dramatically enriched for STRs (Vinces et al. 2009; Sawaya et al. 2013). In coding regions, STRs tend to occur in genes with roles in transcriptional regulation, DNA binding, protein-protein binding, and developmental processes (Persi and Horn 2013; Schaper et al. 2014). These consistent functional enrichments across vastly diverged lineages suggest important functional roles for STRs.

Indeed, analysis of STR variation in the Drosophila Genetic Reference Panel identified dozens of associations between STR variants and quantitative phenotypes in recombinant inbred fly lines (Mackay et al. 2012). Moreover, accumulating evidence from exhaustive genetic studies shows that STR variation has dramatic, often background-dependent phenotypic effects in model organisms (Undurraga et al. 2012b; Michael et al. 2007b; Sawyer 1997; Rosas et al. 2014b;

Scarpino et al. 2013b). Together, these findings suggest that STR variation has the potential to dramatically revise the heritability estimates attributable to genetic factors.

The high STR mutation rate also leads to substantial somatic variation of STR loci within individuals. In fact, this somatic variation, also called microsatellite instability (MSI), has been used for decades as a biomarker for different classes of cancer (Boland et al. 1998). Recent studies demonstrate that organisms exposed to various environmental stresses and perturbations show increased genome instability, including MSI (Kovalchuk et al. 2003; Specchia et al. 2010; Mittelman et al. 2010; Queitsch et al. 2012). MSI may be useful as a biomarker for cellular stress states that may predispose to disease.

The broad interest in STR variation has led to the development of techniques for high-throughput genotyping of STRs (Guilmatre et al. 2013; Duitama et al. 2014) and an explosion of analysis tools for extracting STR variation from existing sequence data (Highnam et al. 2013; Gymrek et al. 2012; Cao et al. 2014). However, the precision of these methods remains limited, due to a combination of low effective coverage of STRs and the lack of robust models for distinguishing technical error from somatic variation. Attempts to use STR variation for GWA in a fashion equivalent to SNV variation may be underpowered and confounded by the unique characteristics of this class of variants. In this review, we discuss the latest advances in these fields, and lay out a set of priorities for the future study of STRs.

STR variation is associated with human genetic diseases

Within coding regions, STR mutations are generally in-frame additions and subtractions of repeat units, resulting in proteins with variable, low-complexity amino acid runs (Gemayel et al. 2010). These mutations can result in phenotypic effects and lead to genetic disorders; several neurological diseases (spinocerebellar ataxias, Huntington's disease, spinobulbar muscular atrophy, dentatorubral-pallidoluysian atrophy, intellectual disability, etc.) are a consequence of dramatically expanded STR alleles (Gatchel and Zoghbi 2005; Fondon et al. 2008; Poeta et al. 2013). Many of these disease-associated STR expansions behave as dominant gain-of-function mutations (Fondon et al. 2008). However, even comparatively modest coding STR variation may confer disease risk or behavioral phenotypes, according to a variety of single-marker association studies (Caspi et al. 2003; Zhang et al. 2012; Eisenegger et al. 2013; Inanir et al. 2013); for instance, variants in separate coding STRs in *RUNX2* are associated with defects in bone

mineralization, higher incidence of fractures (Morrison et al. 2012, 2013); interestingly STR variation in this gene in dogs is also associated with craniofacial phenotypes (Fondon and Garner 2004). Noncoding STR variation in regulatory sequences can affect transcription, RNA stability, and chromatin organization. For instance, certain STR variants alter *CFTR* expression and thus cystic fibrosis status (Eckert and Hile 2009). We take these studies as evidence that STR variation, even in the absence of large expansions, may contribute significantly to the heritability of human traits and genetic diseases.

The severity of the STR expansion-associated diseases may suggest that natural selection should eliminate STRs in functional regions, but several recent studies across many organisms indicate that variable STRs are globally maintained (Li et al. 2012; Mularoni et al. 2007, 2010; Persi and Horn 2013; Schaper et al. 2014). For example, the pre-expansion polyQ-encoding STR in the human gene SCA2 is under positive selection, suggesting that this variable STR is actively maintained in spite of the pathogenic expansions that do occasionally occur and cause spinocerebellar ataxia (Yu et al. 2005). Considering both the evidence of positive selection on STRs and the functional enrichments of STR-containing genes, several authors have proposed that functional STRs are maintained because they confer 'evolvability', or the capacity for fast adaptation (Gemayel et al. 2010, 2012; Vinces et al. 2009; Laidlaw et al. 2007; King 2012). This suggestion is intriguing, in part because many STR mutations are dominant, and, when beneficial, can quickly sweep to fixation. Although we do not further discuss these evolutionary considerations here, they underscore the phenotypic potential of STR variation.

STR variation has dramatic background-dependent effects on phenotype

To date, the functional consequences of unit number variation in selected STRs have been studied in plants, fungi, flies, voles, dogs, and fish (Sawyer 1997; Sureshkumar et al. 2009; Hammock and Young 2005; Undurraga et al. 2012; Rosas et al. 2014; Smukalla et al. 2008), among other organisms. In *Saccharomyces cerevisiae*, STR unit number in the *FLO1* gene accurately predicts the phenotype of cell-cell and cell-substrate adhesion (flocculation); flocculation provides protection against various stresses (Smukalla et al. 2008; Verstrepen et al. 2005). STR variation in yeast promoters has been shown to alter gene expression (Vinces et al. 2009). In *Drosophila melanogaster*, *Neurospora crassa*, and *Arabidopsis thaliana*, natural coding STR variation in circadian clock genes alters diurnal rhythmicity and developmental

timing (Sawyer 1997; Peixoto et al. 1998; Michael et al. 2007; Undurraga et al. 2012). Some have proposed that the large phenotypic responses to selection observed in the Canidae are a consequence of elevated STR mutation rates relative to other mammalian clades (Fondon and Garner 2004; Laidlaw et al. 2007). We can state unambiguously that naturally variable STRs underlie dramatic phenotypic variation in model organisms.

Beyond the observable fact that variable STRs affect phenotype, we can make specific predictions about the components of phenotypic variation that they affect. Both theoretical expectations and empirical data indicate that STR variants are likely to participate in epistatic interactions, and probably more so than most SNVs. One plausible hypothesis is that STRs act as mutational modifiers of other loci, as may be expected intuitively from their elevated mutation rate (**Box 2.1, Fig. 2.I**).

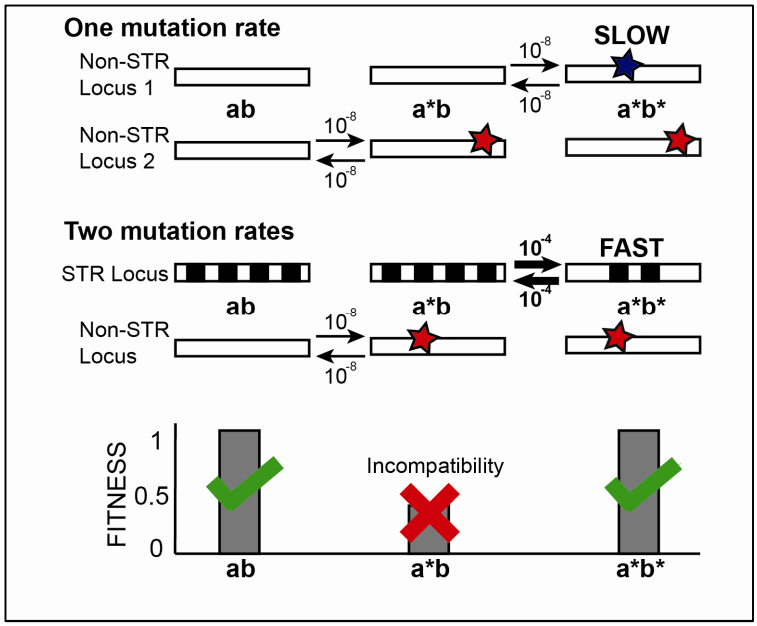

Figure 2.I. A locus with higher mutation rates allows genetic modification of unfavorable genotypes at interacting loci. Top, a model of evolution under epistasis with only one slow mutation rate. Middle, a model of evolution under epistasis with a slow and a fast mutation rate. Boxes represent loci, stars represent SNV-type mutations, black and white checkering indicates an STR locus (*a/b*, *a*/b*, and *a*/b** signify different genotypes). Arrows with numbers represent possible mutations and their respective rates. Bottom, fitness of each genotype under both models. We expect that the model with two mutation rates will occupy the fully derived state (*a*/b**) more quickly.

BOX 2.1: Modifier mutations leading to epistasis are expected in STRs. We have previously proposed that STRs might be more susceptible to genetic interactions [23], as we will briefly explicate here. Consider a simple two-locus haploid model under panmixis, in which loci *A* and *B* each start with a single allele (*ab*) and have the same probability *p* per generation of mutating to a second allele (*a** or *b**), with *p* also as the probability per generation of reverting mutations (Figure I). Let us further assume that *A* and *B* are in sign epistasis [58] (that is, *a*b* and/or *ab** have fitness less than *ab* and *a*b**). To escape the unfavorable *a*b* genotype, the organism may either revert to *ab* or mutate forward to *a*b**. When the *A* and *B* loci have equal mutation rates, we expect that the reversion of a single mutant is just as likely as a second mutation, and consequently that *a*b** individuals will appear only relatively rarely and slowly. However, consider a similar model, in which locus B has an elevated mutation rate $p_b > p_a$. In this case, the *a*b* genotype has a higher probability of a second, modifying mutation to *a*b** than of a reversion to *ab*. Moreover, flux along the other mutational path (*ab* → *ab** → *a*b**) will be increased. In sum, *a*b** genotypes will arise at higher rates, and will attain their equilibrium frequency much more rapidly, if either A or B has an elevated mutation rate [59] (p.131). This scenario can lead quickly to an equilibrium population in which incompatible epistatic alleles are frequent, even though recombinants have lower fitness. Relaxing the assumption of no population structure will further speed this process. Consequently, we would expect STRs and other loci with high mutation rates to be more likely to modify other alleles than loci with lower mutation rates, as long as we assume that all loci are equally capable of genetic interactions. This process may be referred to as 'coadaptation'. For a rigorous model of the evolution of hybrid incompatibility, see Orr [57].

This expectation is borne out in the handful of studies reporting exhaustive genetic analysis of STRs. For instance, in the *Xiphophorus* genus of fish, a genetic incompatibility has recently been attributed to the interaction between the *xmrk* oncogene and an STR in the promoter of the tumor suppressor *cdkn2a/b* (Butler et al. 2007; Scarpino et al. 2013). If the *xmrk* gene product is not properly regulated by *cdkn2a/b*, fish develop fatal melanomas, a two-locus Bateson-Dobzhansky-Muller incompatibility described in classic genetic experiments (**Fig. 2.1A**) (Gordon 1927; Kosswig 1928; Meierjohann and Schartl 2006). Expansions in the *cdkn2a/b* promoter STR are associated with the presence of a functional copy of the *xmrk* oncogene across species, and are thought to functionally repress the activity of the *xmrk* gene product through increased dosage of the tumor suppressor (Scarpino et al. 2013).

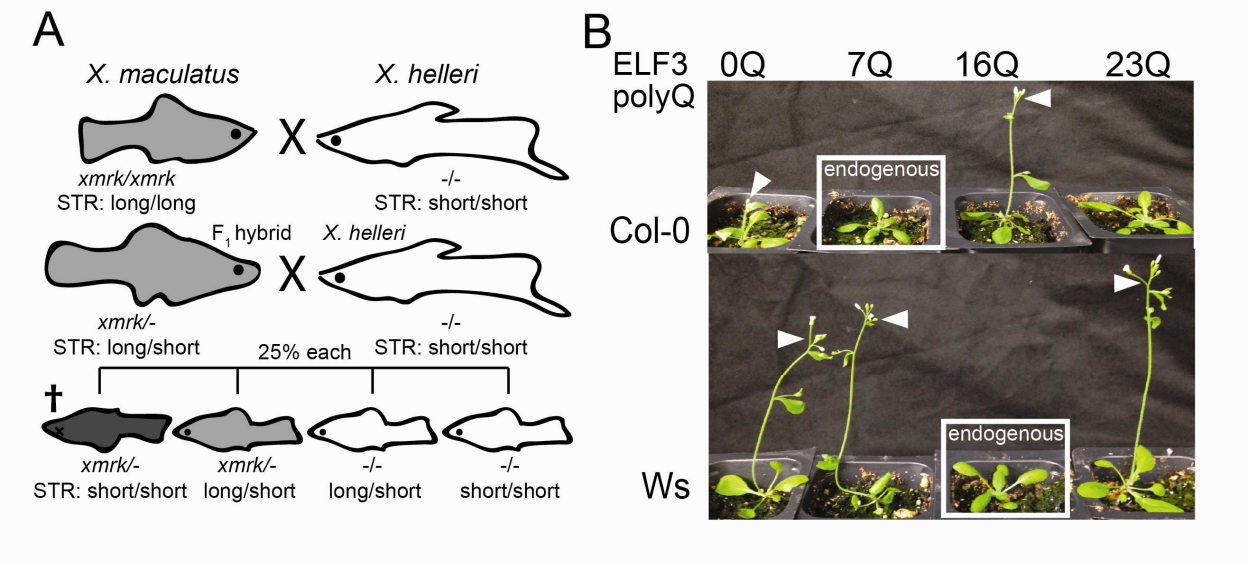

Figure 2.1. Genetic and transgenic analysis reveals STR-mediated incompatibilities. A) the Gordon-Kosswig-Anders cross shows a genetic incompatibility between two fish species in the *Xiphophorus* genus. Modified from Meierjohann and Schartl (Meierjohann and Schartl 2006). F_1 hybrids back-crossed to their *X. helleri* parent yield a 3:1 ratio of viability, where the inviables result from co-segregation of the functional *xmrk* gene and a short STR allele in the *cdkn2a/b* promoter. Shading indicates melanism conferred by *xmrk*. **B)** genetic background is epistatic to effects of *ELF3* STR variation in *A. thaliana*. Expression-matched transgenic plants with various alleles of the *ELF3* STR in the Columbia (Col-0) and Wassilewskija (Ws) backgrounds, showing endogenous, exogenous, and synthetic ("0") alleles in each background (Undurraga et al. 2012b). White boxes indicate transgenic plants carrying the *ELF3* STR endogenous to their respective background; white arrowheads indicate early-flowering *ELF3* STR genotypes (*elf3* mutants and poorly-functioning *ELF3* STR alleles confer early flowering).

Similarly, we have shown that natural variation in the polyQ-encoding *ELF3* STR significantly affects all ELF3-dependent phenotypes in the plant *A. thaliana*, with *ELF3* STR length and phenotype showing a strikingly nonlinear relationship (**Fig. 2.1B**)(Undurraga et al. 2012). Some naturally occurring *ELF3* STR variants phenocopy *elf3*-loss-function mutants in a common reference background (**Fig. 2.1B**), suggesting background-specific modifiers. Indeed, when we compare the phenotypic effects of each *ELF3* STR variant between two divergent backgrounds, Columbia (Col-0) and Wassilewskija (Ws), we find dramatic differences. The endogenous STR alleles from these two strains (Col-0 7 units, Ws 16 units) show mutual incompatibility when exchanged between backgrounds. The ELF3 protein is thought to function as an "adaptor protein" or physical bridge in diverse protein complexes (Yu et al. 2008; Nusinow

et al. 2011). We speculated that background-specific polymorphisms in these interacting proteins underlie the *ELF3* STR-dependent background effect.

Also in *A. thaliana*, a variable STR in the promoter of the *CONSTANS* gene has been linked to phenotypic variation in the onset of flowering (Rosas et al. 2014). *CONSTANS* encodes a major regulatory protein that promotes flowering. Transgenic experiments demonstrate that this regulatory STR variation affects *CONSTANS* expression and hence onset of flowering. However, the effects of this STR variation depend on the presence of a functional allele of *FRIGIDA*, a negative regulator of flowering that is highly polymorphic across *A. thaliana* populations.
A dramatic example of incompatibility can be found in an intronic repeat in the *IIL1* gene in *A. thaliana*, which was found to be dramatically expanded in one strain (Sureshkumar et al. 2009). The expansion delayed flowering under high temperatures, but when crossed into the reference genetic background, a strongly interacting locus modifies this phenotype.

In the *Drosophila* genus, coding STR variation in the *per* gene co-evolves with other variants (Peixoto et al. 1998, 1993). Transgenic flies expressing chimeric *per* genes with a *D. melanogaster* STR domain fused to a *D. pseudoobscura* flanking region (and vice versa) have arrhythmic circadian clocks, indicating the modifying effect of flanking variation in generating an STR-based genetic incompatibility. Among STRs subjected to exhaustive genetic study, to our knowledge, only the yeast *FLO1* coding STR has no known modifiers due to variation in genetic background (Smukalla et al. 2008).

In addition to these exhaustive genetic studies, there are several other observations that support the role of the genetic background in controlling the phenotypic effects of STRs. For instance, experiments in *Caenorhabditis elegans* and human cells indicate that the phenotypic effects of proteins with expanded polyQ tracts are modulated by genetic background (Gidalevitz et al. 2013), or by variants in interacting proteins (Metzger et al. 2008). In humans, genetic association studies indicate the existence of genetic modifiers of polyQ expansion disorders for both Huntington's disease (Rubinsztein et al. 1997) and spinocerebellar ataxias (Zühlke et al. 2005). Taken together, these experimental and observational data support our argument that functional STRs are likely to be enriched for variants in epistasis with other loci.

STRs with background-dependent phenotypic effects tend to either encode polyQ tracts or reside in promoter regions. There are good reasons to expect that these STR classes might be enriched in DNA/protein-protein interactions that could underlie epistasis. PolyQ tracts,

specifically, often bind DNA surfaces (Escher et al. 2000), and an analysis of human protein interactome data found that polyQ-containing proteins engage in more physical interactions with other proteins than those without polyQs (Schaefer et al. 2012). Similarly, noncoding STRs in regulatory regions may compensate for mutations in trans-acting factors, as observed for the STRs in the *cdkn2a/b* promoter in *Xiphophorus* (Scarpino et al. 2013) and in the *CONSTANS* promoter in *A. thaliana* (Rosas et al. 2014). We suggest that polymorphisms in protein interaction partners or in transcriptional regulators are plausible explanations for the observed background effects. In summary, we expect that STR variation is likely to contribute a substantial epistatic component to heritability, which has important implications for their use in explaining phenotypic variation.

Analytical tools and genotyping methods continue to struggle with STR-specific challenges.

To fulfill the promise of STR variation for explaining heritability, we need accurate, genome-wide assessment of STR variation in populations of humans and other organisms. The scientific community has tackled this problem in a flurry of recent studies describing methods for genotyping STRs genome-wide (**Table 1**). Specifically, in the last two years, several analytical tools have been developed to call STR genotypes from whole-genome-sequencing data (Gymrek et al. 2012; Highnam et al. 2013; Cao et al. 2014). These tools attempt to address the two major challenges for genotyping STRs: poor mappability due to low sequence complexity and high technical error rate due to amplification stutter.

Table 1. Technologies for assessing STR variation by high-throughput sequencing.

Name	Data Source	Analysis strategy	Accepted coverage[a]	Reported accuracy	Reported efficiency	Limitations	Ref.
lobSTR	Human, whole-genome[b,c]	Align to modified reference	1 read	88%-95%	0.2% of reads are informative	Depends on depth of sequencing and length of reads	(Gymrek et al. 2012b)
RepeatSeq	Human, whole-genome[b,d]	Align to reference, locally realigned	2 reads	92%	Not reported	Depends on depth of sequencing and length of reads	(Highnam et al. 2013b)
STRViper	*A. thaliana* whole-genome[b,d]	Compare insert size to reference	10 reads	74%	Not reported	Cannot call STR unit number genotypes	(Cao et al. 2014c)
Array Capture	Human, array capture[b]	RepeatSeq	2 reads	88%-92%	2.2% informative reads	Low enrichment for STR-spanning reads	(Guilmatre et al. 2013b)
SureSelect RNA probe capture	Human, target enrichment, Roche 454	Locally align flanking regions	4 reads	88%-95%	27% informative reads	Expensive probe design, captured only 60% of targeted STRs	(Duitama et al. 2014a)

a: Minimum coverage of a single STR that is considered sufficient to call a genotype.

b: Sequence data from Illumina HiSeq technology.

c: data references: (Green et al. 2010; Reich et al. 2010)

d: data references: (Abecasis et al. 2010, 2012)

e: data references: (Gan et al. 2011b; Cao et al. 2011b)

To accurately map an STR sequence read and retrieve its unit number genotype, the sequence read must span the STR of interest and include some unique flanking sequence. This requirement limits the length of STRs that can be accurately genotyped and decreases effective STR coverage compared to average whole-genome-sequencing coverage (**Fig. 2.2**). For this reason, much of the existing sequencing data, which consists largely of short reads (36 bp, 50 bp, or 76 bp) with only modest genome coverage (5-20X) is not suitable for accurate, genome-wide calls of STR genotypes; only a fraction of STRs, mostly short ones, can be assessed with some confidence (**Fig. 2.2**).

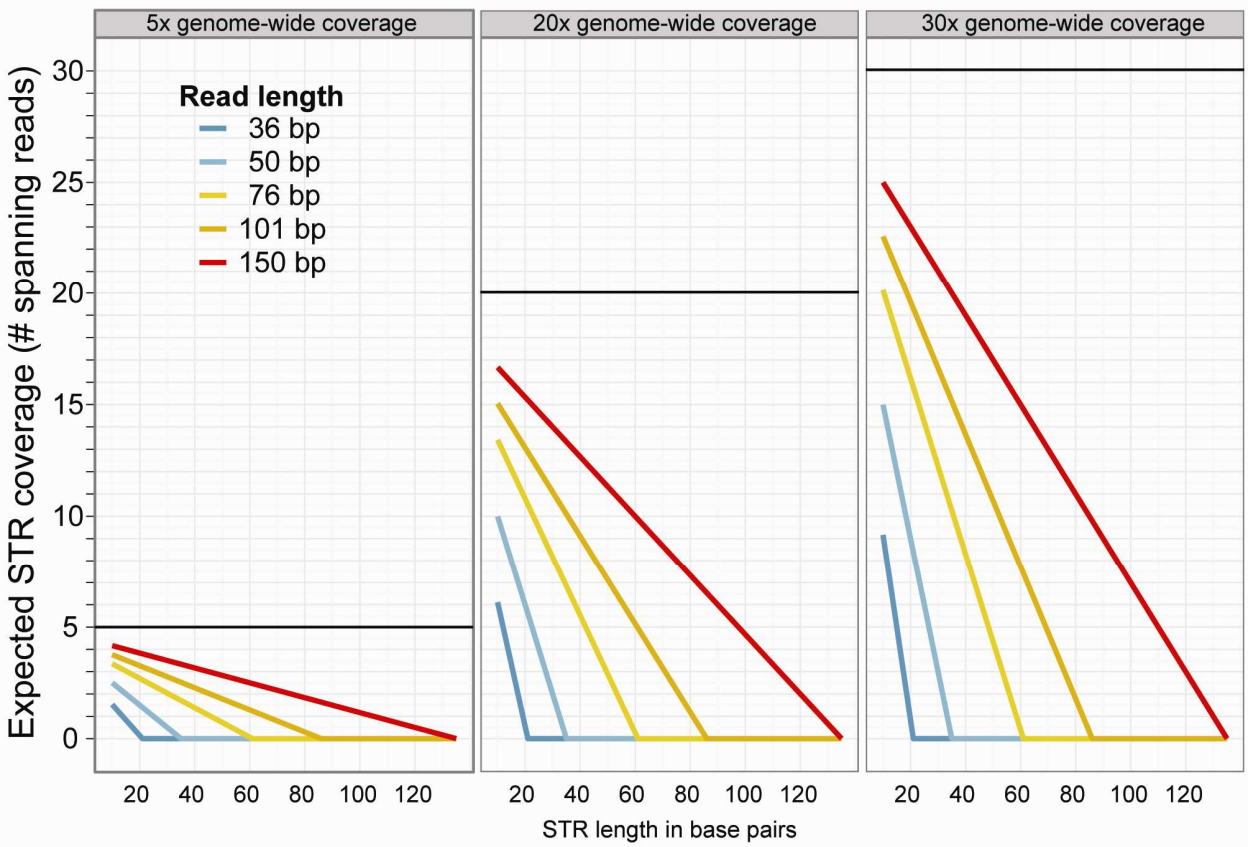

Figure 2.2. Effective reduction in STR coverage in whole-genome sequencing. Expected coverage of STRs for various sequencing depths and read lengths. We assumed 8 bp of flanking sequence on either side (per requirement for LobSTR software (Gymrek et al. 2012b)). Black bars indicate nominal sequencing coverage for each scenario. 4-5X coverage (left panel) is typical for genomes in the human 1000 Genomes Project (Abecasis et al. 2010); 15-20X coverage is typical for genomes in the *A. thaliana* 1001 Genomes Project (Cao et al. 2011b; Gan et al. 2011b).

Moreover, these analytical tools estimate technical error based on STR genotypes from sequenced homozygous or haploid genomes, ignoring somatic alleles within individuals (which are to be expected for STRs even in primary tissues, occurring at rates 10^4-10^5 times higher than SNV somatic mutations) (Slagboom et al. 1991; Golubov et al. 2010; Lee et al. 2010; Chapal-Ilani et al. 2013). In the face of somatic STR variation and high technical error, the probability-based error models of these tools require substantial read coverage to call germ-line STR genotypes with confidence. However, because of the low effective coverage of STR loci (**Fig. 2.2**), STR genotype calls are based on as few as one to two STR-spanning reads (Gymrek et al. 2012b; Highnam et al. 2013b) (**Table 1**). Calls based on so little data may not be accurate even

for homozygous germline alleles. Calling heterozygous STR genotypes remains difficult with the modest coverage of most available whole-genome-sequencing data, such as found in the 1000 Genomes Project (Willems et al. 2014), which becomes even more challenging when potential somatic mutations contribute to a heterogeneous sample population. To illustrate this challenge, consider a heterozygous ~30 bp-STR locus and whole-genome sequencing with 101 bp-reads at 5x coverage – this scenario is likely to yield just three STR-spanning reads (**Fig. 2.2**). These three reads may represent one, two, or three different alleles, representing any mixture of two different germ-line alleles, somatic alleles, or technical error, making an accurate call quite difficult.

Others have attempted to genotype STRs using whole-genome-sequencing data from paired-end reads (50bp) of size-selected genomic fragments (Cao et al. 2014), similar to strategies used to detect large insertions or deletions (Chen et al. 2009; Hajirasouliha et al. 2010; Qi and Zhao 2011; Grimm et al. 2013). This approach is limited by the resolution of gel electrophoresis in the size selection of DNA fragments. Consequently, this method cannot determine STR unit number genotypes, but rather reports whether an STR is variable across samples. The authors argue that this approach is the most accurate for population-level detection of STR variability (Cao et al. 2014), but it is not informative for discerning the relationship between STR unit number genotype and phenotype.

Although these analysis tools represent important and useful advances, their limitations illustrate that ‘dustbin-diving’ of whole-genome-sequencing data may not suffice for accurate population-scale genotyping of STRs genome-wide. Alternative approaches that enrich for STR-spanning sequencing reads are needed. Indeed, two such approaches have been recently published. Both use targeted capture of STRs to enrich for STR-spanning reads combined with high-throughput sequencing compatible with midsize-reads (101 bp, 500 bp) (Guilmatre et al. 2013; Duitama et al. 2014). Targeted STR capture requires the design of STR-specific probes (or rather probes specific to their unique flanking sequences) and involves additional sequencing, but these approaches can dramatically increase the number of informative reads, therefore providing substantial STR coverage for accurate genotyping calls (**Table 1**). For example, the SureSelect-RNA-probe capture method reports 27% informative STR-spanning reads compared to the 0.2 % informative reads found in whole-genome-sequencing data (**Table 1**). This increase in informative reads is a major advantage over whole-genome resequencing because STRs

represent only a small fraction of the genome overall (Guilmatre et al. 2013; Duitama et al. 2014). Although targeted capture combined with high-throughput sequencing appears to be a cost-effective alternative for accurate STR genotyping compared to whole-genome sequencing, distinguishing heterozygous alleles, somatic variants, and technical error remains a challenge. We suggest that recent innovations in single-molecule targeted capture (Hiatt et al. 2013) should be useful in distinguishing these categories and in further increasing enrichment of informative, STR-spanning reads.

Lack of statistical models for detecting STR-phenotype associations in GWA.

Assuming that we obtain accurate, population-scale genotype data for STRs, we may not yet have statistical tools appropriate for detecting STR associations with phenotype (Hannan 2010). In diploid organisms, a biallelic SNV is typically analyzed by modeling phenotype as a function of the number of non-reference alleles at that locus (0, 1, or 2) in each individual. A null hypothesis of no monotonic relationship between phenotype and the allele count is then formulated and tested (Hayes 2013b). This framework cannot accommodate more than two alleles, which we would expect for many STRs. Simply using tagged SNVs linked to STRs to perform GWA is unfeasible, because linkage disequilibrium decays very quickly between SNVs and STRs across human populations (Willems et al. 2014).

To address these complications, a previous study attempted GWA between STR genotypes and human disease phenotypes by comparing relative frequencies of various alleles in pooled DNA from cases and controls (Oka et al. 2003). By pooling samples, this approach eases the analysis of multiallelic loci, but it loses information by ignoring specific individuals.

In a more recent study, the authors used logistic regression and the analysis of variance to detect associations between STR alleles and quantitative phenotypes in an inbred *Drosophila* mapping population (Mackay et al. 2012). Given that significant associations were detected, such approaches may be sufficiently powerful in recombinant inbred lines. However, their strategy relied on homozygosity, and considered multiallelic STRs in a pairwise fashion, so these straightforward methods will lose power with outbred populations and multiallelic STRs.

The central confounder of these studies is that most STRs of appreciable variability (and thus, interest) are multiallelic, as a simple consequence of the STR mutational mechanism (Legendre et al. 2007). This multiallelic feature could be accommodated by treating STR alleles

categorically, but this choice entails a corresponding reduction in power, because many alleles are rare.

Some studies have reported linear associations between STR unit number and quantitative phenotypes (Sawyer 1997; Smukalla et al. 2008), suggesting that using simple tests of linear correlation between these variables may be a powerful option. However, this linearity (or even monotonicity) of the relationship between STR unit number genotype and phenotype is a poorly-supported assumption (Undurraga et al. 2012). Nonetheless, STR unit number is a numerical variable, and it would be preferable to gain power from treating it as such. For instance, more similar STR unit number genotypes might be associated with more similar phenotypes, but this intuition may be difficult to generalize.

Lastly, both intuition (**Box 1**) and the studies discussed above lead us to expect that relatively many phenotypically relevant variable STRs will show epistasis with other loci. This epistasis will reduce power in tests of association between STRs and phenotype (Mackay 2014), given the inadequacy of the current paradigm of quantitative genetics in detecting and modeling the effects of epistasis (Nelson et al. 2013; Mackay 2014). At present, targeted and exhaustive genetic studies (as described above) are the only effective method for understanding the effects of epistasis.

In total, these obstacles present a daunting challenge for the integration of STR genotypes into the current genotype-phenotype maps. Overall, we call for a reappraisal of statistical methodologies for use in GWA with STR variation to account for these various STR-specific confounders.

Somatic STR variation may be a sensitive marker for increased disease susceptibility.

It has been appreciated for some time that the high STR mutation rate leads to somatic variation within individuals in addition to germ-line variation between individuals (Armour et al. 1989; Slagboom et al. 1991). This somatic STR variation is particularly noticeable in tumor tissues, but is also measurable in primary tissues (Armour et al. 1989; Slagboom et al. 1991). While these findings immediately led to systems of classification for tumor types and clones (Lee et al. 2001; Kim et al. 2013; Chapal-Ilani et al. 2013), the investigation of somatic STR variation (or MSI) may also inform us about general phenotypic states and disease susceptibility.

Patients with various complex diseases tend to carry a greater load of rare germ-line variants than unaffected control groups (Heng 2010). It is widely assumed that these rare variants contribute in some fashion to these disorders (Girirajan et al. 2011c); however, an alternative interpretation holds that they are signs of stochastic genome instability, which when increased leads to higher susceptibility to complex diseases. (Heng 2010). Increased genome instability will increase somatic variation, which may then serve as a read-out of disease susceptibility (Heng 2010).This alternative interpretation has some support from empirical data. For instance, perturbation of the molecular chaperone Hsp90, which stabilizes diverse DNA repair proteins, leads to increased somatic STR mutation rates in human cells; in various model organisms Hsp90 perturbation increases transposon mobility and intrachromosomal homologous recombination (Mittelman et al. 2010; Queitsch et al. 2012; Kovalchuk et al. 2003; Specchia et al. 2010). Hsp90 perturbation also increases the penetrance of many genetic variants in flies, plants, fish, worms and yeast, suggesting that increased genome instability and increased phenotypic heritability are associated (Queitsch et al. 2012). If this association also applies to disease phenotypes, increased genome instability may predict higher disease susceptibility.

Consequently, although somatic MSI may not be the cause of disease phenotypes, it may serve as a biomarker for individuals who are more vulnerable to environmental and genetic perturbations leading to disease. Again, this strategy hinges on the development of cost-effective technologies for screening panels of STRs for somatic mutations across many humans, which will require specific strategies.

Another possibility is that somatic variation is itself phenotypically relevant, or even plays a role in developmental processes. It is known that STRs are enriched in genes with neuronal function (Bolton et al. 2013); some have even proposed that such somatic mutation is a component of normal neuronal development in humans (Nithianantharajah and Hannan 2007). If this is the case, then a greater appreciation of somatic variation will be necessary to understand canonical developmental processes. Collectively, STR variation within (in addition to between) individuals has great potential as a read-out for disease susceptibility, and perhaps also as a cause of phenotypic variation itself.

Concluding remarks

The study of STRs and other under-ascertained genomic elements has the potential to reshape our model of the heritability of complex diseases and traits, both in terms of the overall proportion of heritability explained, and in terms of the components of heritability themselves (Outstanding Questions). Experimental studies in model organisms have taught us that the phenotypic effects of genome-wide STR variation are both dramatic and impossible to understand without taking epistasis into account. In the future, our understanding will be improved by 1) accurate STR population-scale and somatic genotyping, 2) more appropriate statistical methods for analyzing STR-phenotype associations, and 3) a broader description of epistasis between STR variation and other loci in determining phenotype.

OUTSTANDING QUESTIONS

- **In light of wide-spread epistasis, what statistical and experimental tools can quantify the effect of STR variation on phenotype?**
- **Can inexpensive, accurate tools be developed for germ-line and somatic STR genotyping?**
- **Will somatic STR variation be effective as a readout for disease susceptibility?**

GLOSSARY

Short tandem repeat (STR): a repetitive nucleotide sequence that consists of many copies of a short sequence in tandem (ex. CAGCAGCAGCAG). STRs are frequently called **microsatellites**.
Single nucleotide variant (SNV): Variant that consists of a change at a single nucleotide position. Common SNVs are sometimes called single nucleotide polymorphisms (SNPs).
Heritability: The fraction of variation in a phenotype across a population that can be attributed to genetic differences.
Epistasis: Non-reciprocal interactions of non-allelic gene variants, due for instance to functional interdependence between gene products in a protein complex or metabolic pathway.
Genome-wide association (GWA): A set of methods by which each of a large number of genetic variants genome-wide is tested for statistical associations with a phenotype. Often referred to in the context of **genome-wide association studies** (**GWAS**).

Complex disease, complex traits: Complex diseases or traits are phenotypic characters thought to be affected by multiple genetic and environmental factors.

Somatic variation: Genetic variation across somatic cells or tissues of an organism, which are generally not inherited by offspring (which inherits instead **germ-line variation**). Generally arises from mutations in specific cell lineages after early development.

Microsatellite instability (MSI): Somatic variation of STRs (microsatellites) associated with phenotypic changes such as cancer, often due to mutations in DNA repair genes.

Bateson-Dobzhansky-Muller incompatibility: Hybrid incompatibilities observed when crossing two close species or divergent strains of a species against one another. Caused by the segregation of non-parental allele combinations to individuals, resulting in a dysfunctional genetic interaction (negative epistasis).

Acknowledgements

This work was supported by grants from the National Human Genome Research Institute Interdisciplinary Training in Genomic Sciences (2T32HG35-16 to MOP, T32 HG00035 to KDC) and the National Institute of Health New Innovator Award (DP2OD008371 to CQ). We are grateful to members of the Queitsch and Shendure laboratories for helpful discussions, and to Wen Huang and David Mittelman for responding to email inquiries.

CHAPTER THREE

MIPSTR: A METHOD FOR MULTIPLEX GENOTYPING OF GERM-LINE AND SOMATIC STR VARIATION ACROSS MANY INDIVIDUALS[2]

Keywords: short tandem repeat (STR), microsatellite instability (MSI), genetic heterogeneity, somatic variation, STR variation

Abstract

Short tandem repeats (STRs) are highly mutable genetic elements that often reside in functional genomic regions. The cumulative evidence of genetic studies on individual STRs suggests that STR variation profoundly affects phenotype and contributes to trait heritability. Despite recent advances in sequencing technology, STR variation has remained largely inaccessible across many individuals compared to single nucleotide variation or copy number variation. STR genotyping with short-read sequence data is confounded by (1) the difficulty of uniquely mapping short, low-complexity reads and (2) the high rate of STR amplification stutter. Here, we present MIPSTR, a robust, scalable, and affordable method that addresses these challenges. MIPSTR uses targeted capture of STR loci by single-molecule Molecular Inversion Probes (smMIPs) and a unique mapping strategy. Targeted capture and mapping strategy resolve the first challenge; the use of single molecule information resolves the second challenge. Unlike previous methods, MIPSTR is capable of distinguishing technical error due to amplification stutter from somatic STR mutations. In proof-of-principle experiments, we use MIPSTR to determine germ-line STR genotypes for 102 STR loci with high accuracy across diverse populations of the plant *A. thaliana.* We show that putatively functional STRs may be identified by deviation from predicted STR variation and by association with quantitative phenotypes. Employing DNA mixing experiments and a mutant deficient in DNA repair, we demonstrate that MIPSTR can detect low-frequency somatic STR variants. MIPSTR is applicable to any organism with a high-quality reference genome and is scalable to genotyping many thousands of STR loci in thousands of individuals.

---

[2] This chapter is under review at Genome Research as "MIPSTR: a method for multiplex genotyping of germ-line and somatic STR variation across many individuals" by KD Carlson, PH Sudmant, MO Press, EE Eichler, J Shendure, C Queitsch.

Introduction

Variation in short tandem repeats (STRs), which are also known as microsatellites, significantly contributes to phenotypic variation, evolutionary adaptation, and human disease (Gemayel et al. 2012). STRs consist of short (2-10 bp) DNA sequences (units) that are repeated head to tail. The presence of multiple identical or nearly identical adjacent sequence units causes frequent errors in recombination and replication, resulting in loss or gain of units. Consequently, STR mutation rates are 10-10,000 times higher than mutation rates of non-repetitive loci (Eckert and Hile 2009; Legendre et al. 2007).

In spite of their hyper-variability, STRs frequently reside in functional DNA, including coding and regulatory regions. STRs are estimated to be present in six percent of human coding regions (Mularoni et al. 2006; O'Dushlaine et al. 2005), highlighting the potential of STR variation to affect disease risk and other complex traits. Coding STRs that vary among humans tend to reside in genes affecting transcription and neural development (Molla et al. 2009). Several severe genetic diseases, including the trinucleotide expansion disorders Huntington's and Spinocerebellar Ataxias (SCA), are a consequence of extended STR alleles that act as dominant mutations (Gatchel and Zoghbi 2005). The severity of STR expansion disorders would suggest that natural selection should remove STRs from functional genomic regions, but some, for example the pre-expansion STR allele in SCA2, are maintained by selection (Yu et al. 2005).

Model organism studies have demonstrated significant functional consequences of even subtle unit number variation in select STRs in plants, fungi, flies, voles, dogs, and fish, among other organisms (Fondon and Garner 2004; Hammock and Young 2005; Michael et al. 2007; Rosas et al. 2014; Sawyer et al. 1997; Scarpino et al. 2013; Undurraga et al. 2012). Similarly to humans, STR-containing genes in these organisms tend to be regulatory genes functioning in transcription, development, and sensing environmental factors (Fondon and Garner 2004; Verstrepen et al. 2005). Adding or subtracting a single STR unit can have dramatic phenotypic effects, such as in the polyglutamine-encoding STR in the circadian clock gene *ELF3* in *Arabidopsis thaliana* (Undurraga et al. 2012). STR unit number can show striking non-linear relationships with phenotype, which may in part be due to extensive epistatic interactions with other loci (Butler et al. 2007; Peixoto et al. 1998; Undurraga et al. 2012). Based on existing evidence, STR variation likely comprises an important component of the genotype-phenotype map (*e.g.,* STRs are a viable explanation for some component of the 'missing heritability' of

genome-wide association studies (Press et al. 2014)), yet due to technological difficulties in genotyping STRs, this component has remained largely undefined.

STRs have almost entirely escaped genome-wide assessment across many individuals due to the complexities of uniquely mapping short, repetitive sequencing reads and the inherently high error rate of STR amplification (*i.e.* amplification stutter). Thus, STR variation is typically excluded or misreported for genomes sequenced with short reads. Recently, several tools have been developed to estimate STR unit number from short read sequencing data (Gymrek et al. 2012; Highnam et al. 2013; Tae et al. 2013). These tools rely on the use of only STR-spanning reads with unique flanking regions to improve mappability and ascertain STR unit number. This restriction imposes size limits (read lengths in extant data are generally 101 bp or less) and greatly reduces coverage of informative reads (**Supplemental Fig. 3.1**). For example, when assessing the genotype of an STR locus of ~30 bp for a genome sequenced with 101 bp reads at 5X coverage, one will have to rely on fewer than three STR-spanning reads on average. Moreover, these tools model technical error due to amplification stutter based on STR genotypes from sequenced homozygous or haploid genomes, ignoring the expected diversity of somatic alleles within individuals. These probabilistic models lose applicability in practice, because STR genotype calls are made with as few as one or two STR-spanning reads. Another recent method uses paired-end sequencing reads to infer variation at STR loci, similar to previous methods to detect large insertions and deletions (Chen et al. 2009; Grimm et al. 2013; Hajirasouliha et al. 2010; Qi and Zhao 2011). Due to the resolution limits of gel size selection, this method infers only whether STRs are variable rather than calling STR unit number genotypes (Cao et al. 2014). Thus, the comprehensive assessment of accurate STR genotypes from short-read sequencing data has remained a largely intractable problem.

Vast numbers of genomes, including genomes of hundreds *A. thaliana* strains have been generated with 36 to 64 bp read lengths (Cao et al. 2011; Gan et al. 2011) that are too short for the aforementioned tools. The existing read lengths and coverage depths of these genomes are sufficient to call most single nucleotide variants (SNVs), but insufficient to understand STR variation. It would be inefficient and costly to re-sequence whole genomes of hundreds of individuals or strains with sufficient depth and the longer reads necessary to understand STR variation (~150-300 bp, >30x coverage) when STRs only make up a small portion of the genome.

The challenges of STR genotyping can be addressed by targeted STR capture to increase the number of STR-spanning reads combined with a sequencing technology that accommodates longer reads to improve mappability and STR genotype calling. Such strategies were recently applied to the human genome, using STR-targeted microarray capture or RNA probe capture prior to sequencing (Duitama et al. 2014; Guilmatre et al. 2013). However, these STR capture methods produced only limited enrichment for STR-containing reads with flanking sequence (2.2% of mappable reads (Guilmatre et al. 2013) and 25% of mappable reads (Duitama et al. 2014) and only marginally improved STR coverage for unit number calls (**Table 3.1**).

Table 1. Technologies for assessing STR variation by targeted capture and high-throughput sequencing.

Name	Sequencing and analysis strategy	Accepted coverage[a]	Reported accuracy	Reads mapped to STR targets	Efficiency of mapped reads	STR targets successfully genotyped	Ref.
Array Capture	Human, Illumina HiSeq, RepeatSeq (Highnam et al. 2013a)	2 reads	88%-92%	38.7%	6.5% informative reads	>= 1 genotype for 54.5% of targets across 8 individuals	(Guilmatre et al. 2013a)
SureSelect RNA probe capture	Human, Roche 454 , locally align flanking regions	4 reads	88%-95%	~60%	40% informative reads	30.1%-36.8% of targets per sample	(Duitama et al. 2014)
MIPSTR-smMIP capture	*A. thaliana*, Illumina MiSeq, map to locus-specific synthetic reference	4 reads	94%-98%	72%	55%-64% informative reads	64% of targets across samples, at least 50% of targets in 90% of samples	

a: Minimum coverage of a single STR that is considered sufficient to call a genotype.

Here, we address the major obstacles of STR genotyping with a robust, scalable, and inexpensive method, MIPSTR. MIPSTR combines STR capture via single-molecule molecular inversion probes (smMIPs) (Hiatt et al. 2013) with mid-size sequencing reads and a unique mapping strategy. In proof-of-principle experiments, we captured and sequenced STRs genome-wide in diverse *A. thaliana* populations, called germ-line STR genotypes with high accuracy, and quantified technical error with single-molecule information. Moreover, enabled by single-molecule degenerate sequence tags, we demonstrate that MIPSTR can capture the same STR locus from thousands of different cells, thereby enabling detection of somatic STR variants with high sensitivity.

Results

Single molecule capture strategy yields highly accurate STR germ-line genotypes

We employed single molecule Molecular Inversion Probes (smMIPs) (Hiatt et al. 2013) to capture STRs, thereby maximizing the number of STR-spanning, informative reads. In a proof-of-principle experiment, we targeted 102 STRs across the genome of the model plant *A. thaliana*, including exonic, intronic, regulatory (AM Sullivan, AA Arsovski, J Lempe, KL Bubb, et al., in press), and intergenic tri- and hexa- nucleotide STRs (**Supplemental Fig. 3.2, Supplemental Table 3.1**). We first applied MIPSTR to the reference *A. thaliana* strain Columbia-0 (Col-0), which has been Sanger-sequenced and for which accurate STR genotypes are available for comparison.

For each targeted STR, we designed a MIP, which is an 80bp oligonucleotide that contains: i) targeting arms which will uniquely hybridize to STR flanking regions, ii) a 12bp degenerate tag to distinguish individual capture events, and iii) a common backbone for PCR and sequencing priming (**Fig. 3.1A**) (Hiatt et al. 2013a). In Col-0, we successfully captured all 102 STR target loci (**Supplemental Fig. 3.3**). After capture, MIPs were amplified for subsequent sequencing. As STR amplification is prone to PCR stutter and rampant technical error, we performed optimizations including modifying amplification conditions, specifically adjusting extension time, extension temperature, and polymerases used (see Methods).

MIPSTR libraries were sequenced using 250bp forward reads paired with 50bp reverse reads on the Illumina MiSeq platform. The 250bp forward reads spanned the ~20 bp ligation targeting arm followed by 200 bp of target sequence (STR sequence and unique flanking sequence) and ~20 bp extension targeting arm (large STR expansions will be missing some or all of the extension targeting arm). MIPSTR can assess STRs up to ~180 bp in length, considerably longer than the STRs currently assessed from whole-genome-sequencing data. The 50 bp reverse reads spanned the 12 bp degenerate tag, which identifies each specific MIP molecule, and the

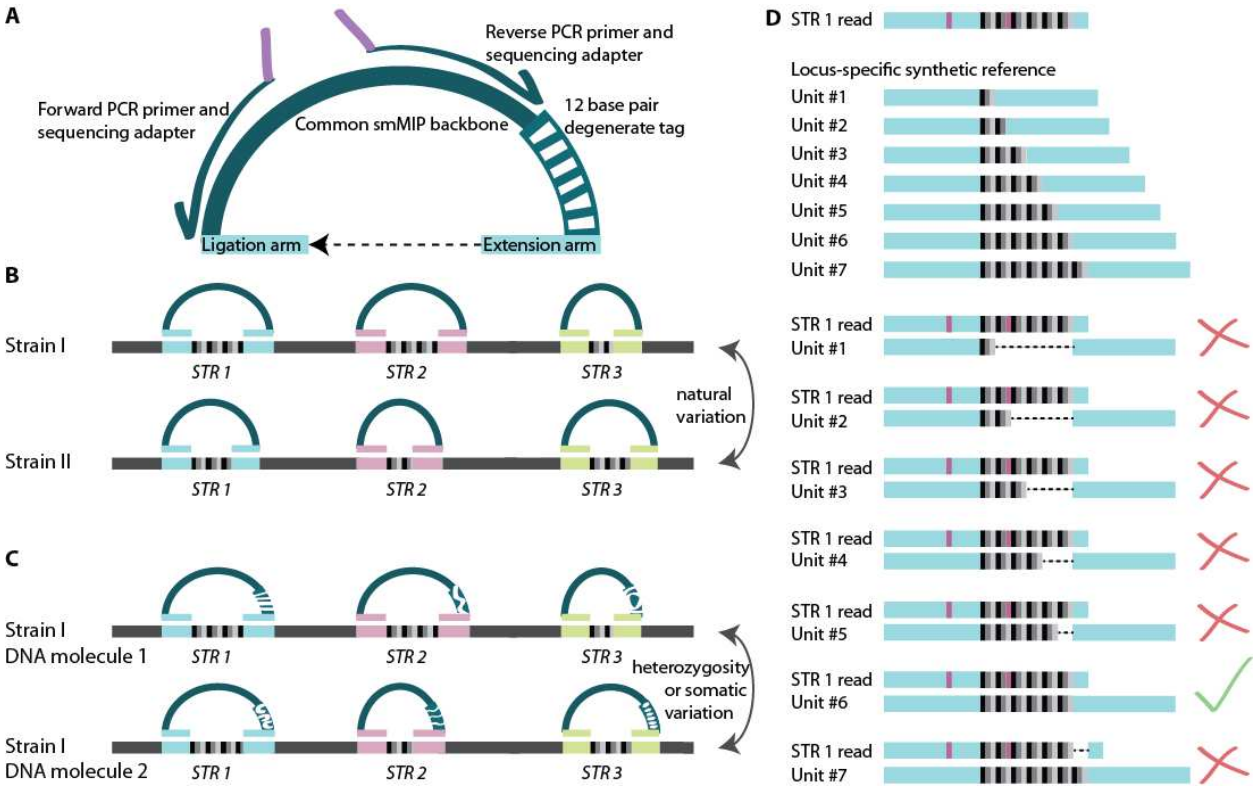

Figure 3.1. MIPSTR determines germ-line and somatic STR variation with a combination of targeted capture, sequencing, and a novel mapping strategy. A) single-molecule molecular inversion probe (smMIP) with common backbone for PCR primer binding (dark-green, also shown PCR and sequencing primers with arrows and purple sequencing adapter), 12 base pair degenerate tag (striped, green/white), and targeting arms with locus-specific, STR-flanking sequence (blue). As shown, one targeting arm is the primer for polymerase extension (extension arm), ligation closes the circle at the other targeting arm (ligation arm). **B**) Applying across individuals identifies germ-line STR variation across genetically diverse individuals. **C**) Applying MIPSTR distinguishes somatic STR variation from technical error, using many degenerate tags (see in **A**). STR variation within a tag-defined read group (*i.e.* reads with the same degenerate tag) is considered technical error. STR variation across tag-defined read groups is considered somatic variation. **D**) MIPSTR maps reads from a given STR locus (based on targeting arm sequence) to locus-specific synthetic references with unit number 1 through 100 (1 through 7 shown here). SNVs (in pink), even if occurring in the STR sequence, do not affect mapping or STR unit number genotype calls.

extension targeting arm (**Fig. 3.1A**). This experimental design allows MIPSTR to omit the computationally costly and error-prone step of mapping repetitive reads of low complexity to whole genomes. We sorted reads according to their MIP targeting arms, and for each MIP, used BWA (Li and Durbin 2009) to map its corresponding reads to a set of synthetic reference

sequences designed specifically for each targeted STR (**Fig. 3.1D**). These synthetic references consisted of the STR sequence from the Col-0 reference genome with all possible STR unit number alleles between 1 and 100, which suffices for STR alleles within our size range. We successfully mapped 72% of all sequencing reads to the targeted loci (**Table 3.1**).

We called a genotype for each mapped read according to the quality of its alignment to an STR allele sequence (BWA alignment scores >= 180 were called as genotypes). Due to our mapping strategy, variation outside of the STR or SNVs within the STR does not affect STR unit number genotype calls (**Fig. 3.1D**). For Col-0, 55% of our mappable reads yielded informative STR unit number calls. Relative to previously described methods, this result represents a dramatic improvement in the number of informative reads per unit of sequencing effort (**Table 3.1**), such that it represents a substantial improvement in the efficiency and accuracy of STR genotyping. We required at least four STR-spanning reads at each locus to call an STR genotype. Ultimately, we called unit number genotypes for 96 out of the 102 examined STR target loci. For these loci, our calls were 96% concordant with the Col-0 reference allele, including the highly variable coding STR in the gene *ELF3* (**Fig. 3.2**) (Undurraga et al. 2012).

Most importantly, unlike any previous method that we are aware of, each STR is represented by many independent capture events of STR loci at the pre-amplification stage. Although amplification introduces technical error, MIPSTR distinguishes between technical error, heterozygosity, and somatic mutations by comparing reads within and between capture events **(Fig. 1C**). The assessment of independent capture events is enabled by the use of smMIPs with degenerate tags (Hiatt et al. 2013), *i.e.* the same STR locus from many different cells is captured by many differently tagged MIP molecules. For each tag-defined read group (*i.e.* reads

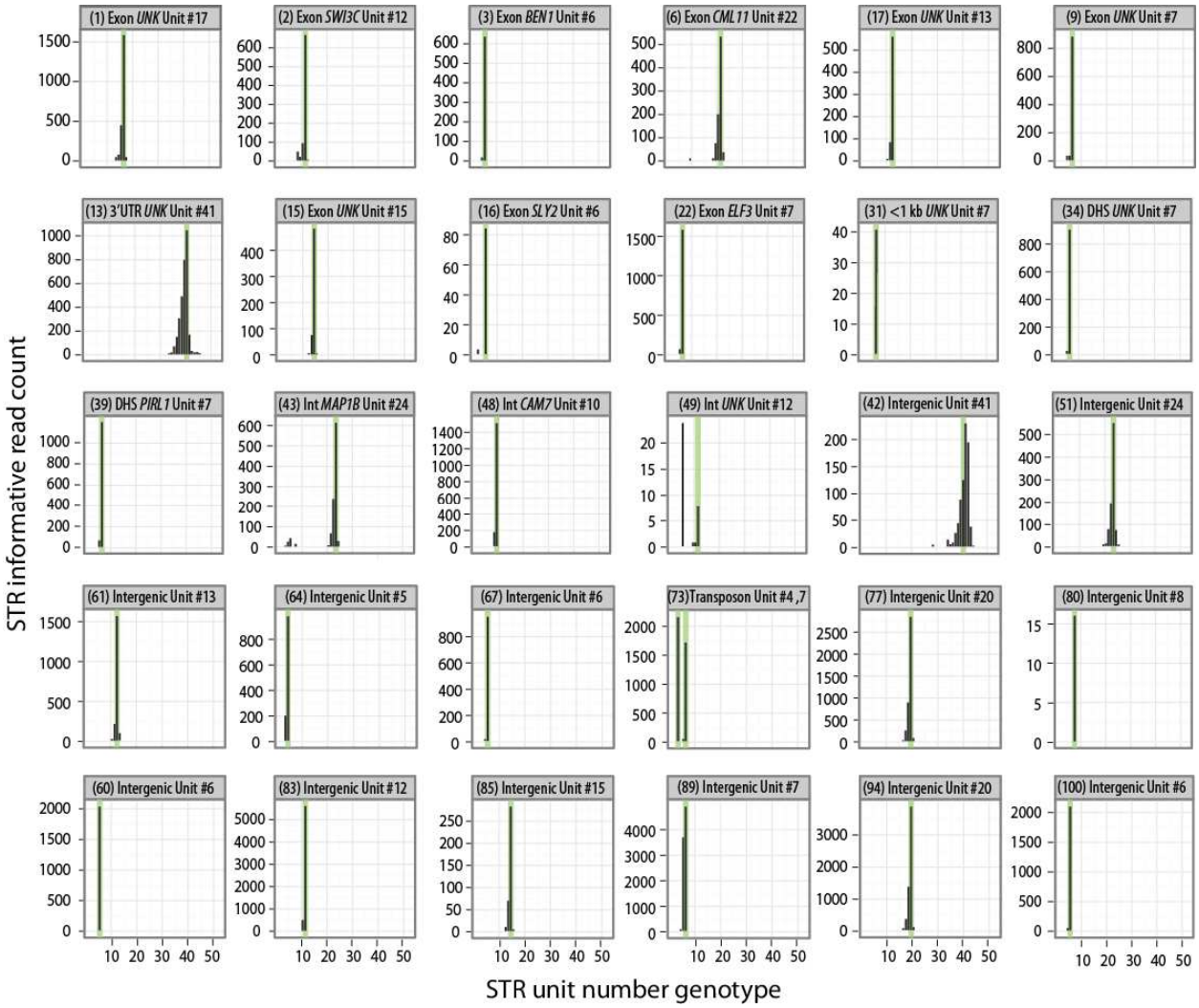

Figure 3.2. MIPSTR accurately determined germ-line STR unit number in the reference strain Col-0. Raw read counts at 30 representative STR loci, with reference genome STR unit number indicated in green. *UNK* indicates gene of unknown function. Numbers shown in parentheses refer to STR IDs (see Supplemental Table 1). Two instances of genomic duplication (residing in transposons) are shown (STR ID 73 and 89) – both alleles showed comparable read count. Note that erroneous calls show low read count or high technical error.

containing the same degenerate MIP tag), we assumed that the mode of called unit numbers across reads is the true allele for this capture event (**Fig. 3**). STR unit number variation within a tag-defined read group is considered technical error (**Fig. 3, Supplemental Table 1**). However, unit number variation observed among different MIP molecules, each representing independent capture events, is potentially the result of heterozygosity, somatic variation, or duplication (**Figs. 1C, 3**). Using the additional information of tag-defined read groups resolves the distribution of

total read counts (compare **Fig. 3A** to **3B, C**) and greatly improves confidence in STR genotype calls. Using information from tag-defined read groups also identified STR loci with consistently high technical error (**Fig. 3**, middle panel, **Supplemental Table 1**), which can be excluded in subsequent analyses. Furthermore, using information from tag-defined read groups has the potential to detect multiple STR alleles within a single individual (**Fig. 3**, right panel).

A. thaliana is an inbreeding plant and hence assumed to be homozygous at the vast majority of loci. Therefore, to test the potential of our method to detect multiple high-frequency alleles of the same STR, we assessed two STR loci present in two nearly identical copies on two different chromosomes in the Col-0 reference genome. For both STRs, the two genomic copies have different STR unit number genotypes in addition to SNV variation, enabling us to readily distinguish them. Indeed, for both STRs, we detected both unit numbers at high levels.

Specifically, for the STR (STR ID 73a and b) with only one SNV difference between duplicate copies, we observed near equal representation of both alleles (**Fig. 3,** right panel). We also observed two tag-defined read groups supporting unit number six, which may represent a somatic STR variant in this individual. Without differentiating tag-defined read groups, reads representing this STR genotype would be interpreted as technical error, like the few reads representing ELF3 STR unit number as six (compare **Fig. 3** left panel to right panel). This example demonstrates the importance of including single-molecule information in STR genotype analysis.

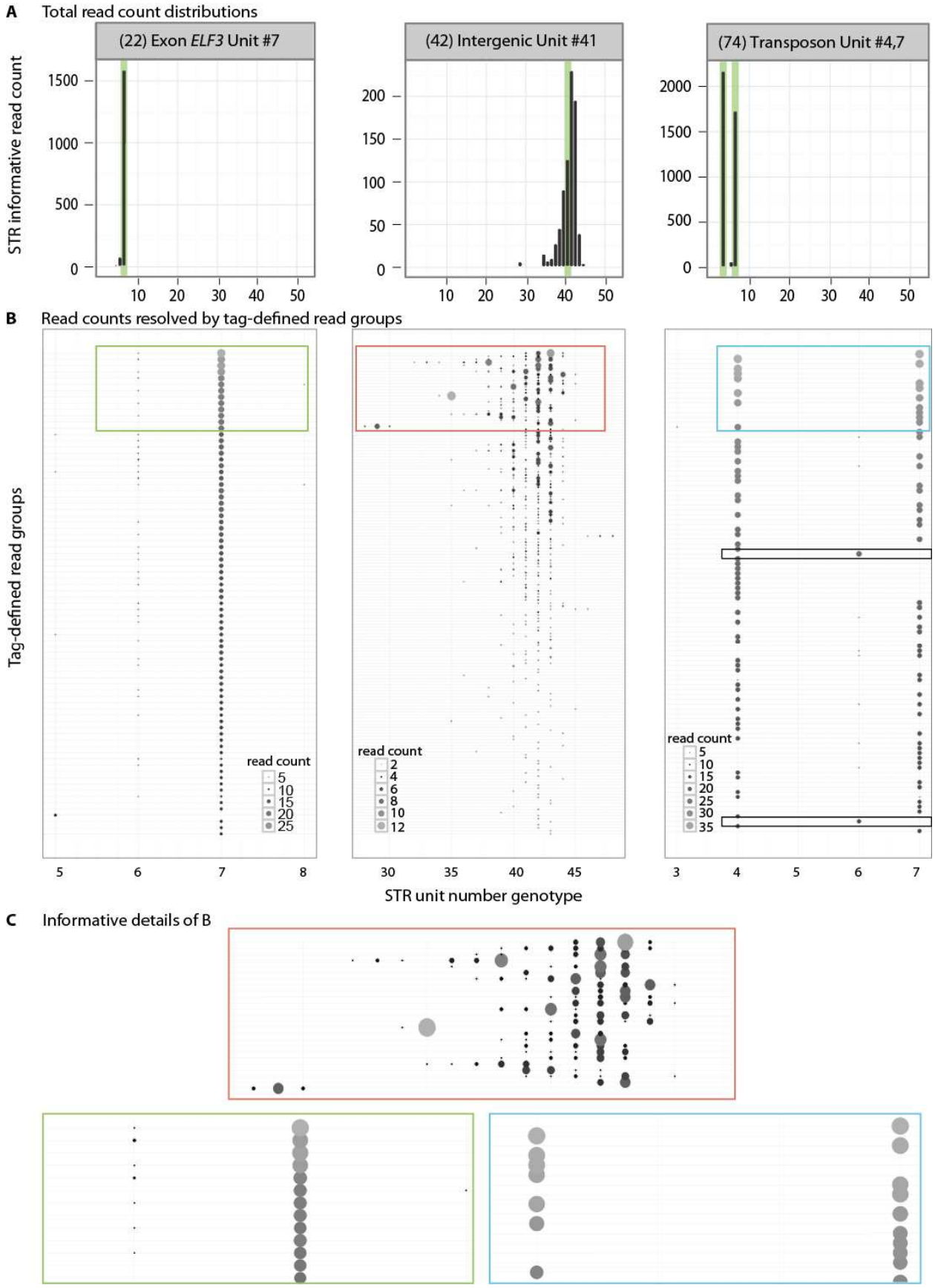
A
Total read count distributions
(22) Exon ELF3 Unit #7
(42) Intergenic Unit #41
(74) Transposon Unit #4,7
STR informative read count
B
Read counts resolved by tag-defined read groups
Tag-defined read groups
read count
STR unit number genotype
C
Informative details of B

Figure 3.3. MIPSTR distinguished technical error from somatic variation. A) Three histograms from Figure 2 with total read counts. Left, the known *ELF3*-STR unit number is clearly supported by the modal unit number. Middle, this intergenic STR showed great variation in STR unit number; the mode did not support the known STR unit number. Right, this STR resides in two copies in two different genomic locations (transposons). Both known alleles were identified, yet total read counts alone cannot distinguish genomic duplicates from technical or somatic error. **B**) Reads are separated into tag-defined read groups with dot sizes and color representing read count (different scales for each locus, see inset). Colored boxes are shown in detail in C. Left, all tag-defined read groups with one exception supported the known STR unit number seven. Most tag-defined read groups showed low levels of technical error, primarily reads with unit number six (-1), but also five and eight. Middle, separating reads into tag-defined read groups illustrates the extremely high technical error for this STR. The mode of a tag-defined read group was often supported by less than 50% of total reads. Some tag-defined read groups contained as many as six different STR genotypes. We exclude such loci from the analysis of somatic STR variation. Right, as expected for a duplicate STR or a heterozygote, approximately half of the tag-defined read groups support each of the known STR genotypes with very little technical error. We also observed evidence of a somatic STR allele with unit number six, which was supported by two tag-defined read groups (boxed, black outline). Note the absence of either known STR allele for these tag-defined read groups. This STR genotype is also visible in the total read count histogram (**A**, right), where it would be interpreted as technical error by other methods. **C**) Detailed views of plots in B; outline color corresponds to respective plot.

Furthermore, we found evidence for the duplication of an intergenic STR that is located amidst multiple transposons; this duplication is not present in the Col-0 reference assembly (**Fig. 3.2**, STR ID 89). As for the other duplicated STRs, the two alleles, in this case six and seven, were supported by approximately equal number of tag-defined read groups in multiple Col-0 siblings. These results suggest that MIPSTR can readily identify heterozygous and somatic STR variants, which have been largely inaccessible by previous analytical or empirical methods (Gymrek et al. 2012; Willems et al. 2014; Guilmatre et al. 2013; Highnam et al. 2013; Duitama et al. 2014).

MIPSTR accurately determines STR unit number genotypes across diverse *A. thaliana* strains

We applied MIPSTR to 96 genetically diverse strains of *A. thaliana*. These strains have been assessed for over 100 quantitative phenotypes and have been previously sequenced, primarily with 36 to 64 bp reads at a coverage of ~20X, to detect SNVs and structural variation (Cao et al. 2011; Gan et al. 2011). STRs evolve on a different time scale than SNVs, so linkage

disequilibrium between STRs and SNVs breaks down quickly (Willems et al. 2014). Therefore, we cannot use linked SNV data to understand the relationship between STR unit number genotype and phenotype. Given the strong potential of STR variation to cause phenotypic variation, we set out to call STR genotypes across many divergent individuals and to show how even data for only 100 STR loci can improve our understanding of the genotype-phenotype map.

We determined the genotypes of the 100 STRs across the 96 diverse strains of *A. thaliana* including the reference strain Col-0 for a total of 9600 targeted STR loci in one Illumina MiSeq v2 sequencing run. MIPSTR scaled well to this task; both the number of targeted loci and the number of examined genomes can be readily increased by several orders of magnitude. STRs tend to be surrounded by repetitive sequence and AT rich regions, but in spite of this challenge, we successfully captured STR loci genome-wide for these genetically divergent strains. Specifically, we captured at least 50 STR loci in 86 out of 96 strains (90%, **Table 3.1**) and at least 75 STR loci in 59/96 strains (61%).

To apply MIPSTR to multiple strains, we pooled the 96 strain-specific capture libraries, each with a unique strain barcode on the reverse PCR primer, and sequenced as described above. For these pooled libraries, we sorted reads first by strain-specific barcode, then by targeting arm to identify the STR locus and degenerate MIP tag to identify reads originating from the same capture event (**Fig. 3.1B**). Similarly to our results with the reference strain Col-0, we were able to map 72% of sequence reads to their STR target loci and of those 64% were informative for calling STR unit number genotypes (**Table 3.1**). In this experiment, the Col-0 library represented ~1% of the total sequence reads, which should greatly reduce the information for each STR compared to our single Col-0 library run. Despite this dramatic reduction in information content, we could accurately call germ-line STR unit number genotypes for 97% of loci (64 out of 66 loci with at least 4 STR-spanning reads) (**Supplemental Table 3.2**). Comparing MIPSTR calls for the *ELF3*-STR to genotype calls from previous Sanger sequencing (Undurraga et al. 2012), MIPSTR performed with 98% accuracy (51 out of 52 strains) (**Fig. 3.4**). As previously discussed, using information from tag-defined read groups aided us in resolving STR genotypes. For example, for the strain Kin-0, total counts supported unit number 18 and 19 for the *ELF3* STR (**Fig. 3.4**). Resolving read counts by tag-defined read groups enabled us to eliminate technical error and call 19 units as the correct Kin-0 *ELF3* STR unit number. Across all 96 strains, we called STR unit number for 60% or more of STR loci in 62% of strains, with a total

of 6,179 STR unit number genotypes (out of 9,600 targets or about 64% of targets) determined with a single Illumina MiSeq v2 sequencing run. As previously shown, additional sequencing is expected to yield many more capture events and thus more complete coverage across STRs (Turner et al. 2009).

The unit number, unit length, and purity of a given STR locus in a high-quality reference genome predict its variation across individuals (Legendre et al. 2007). STRs with high unit number, short unit length, and high purity are typically highly variable. With population-scale STR genotypes in hand, we addressed how well predicted variation of STRs (VARscore) (Legendre et al. 2007) correlated to observed variation across *A. thaliana* strains.

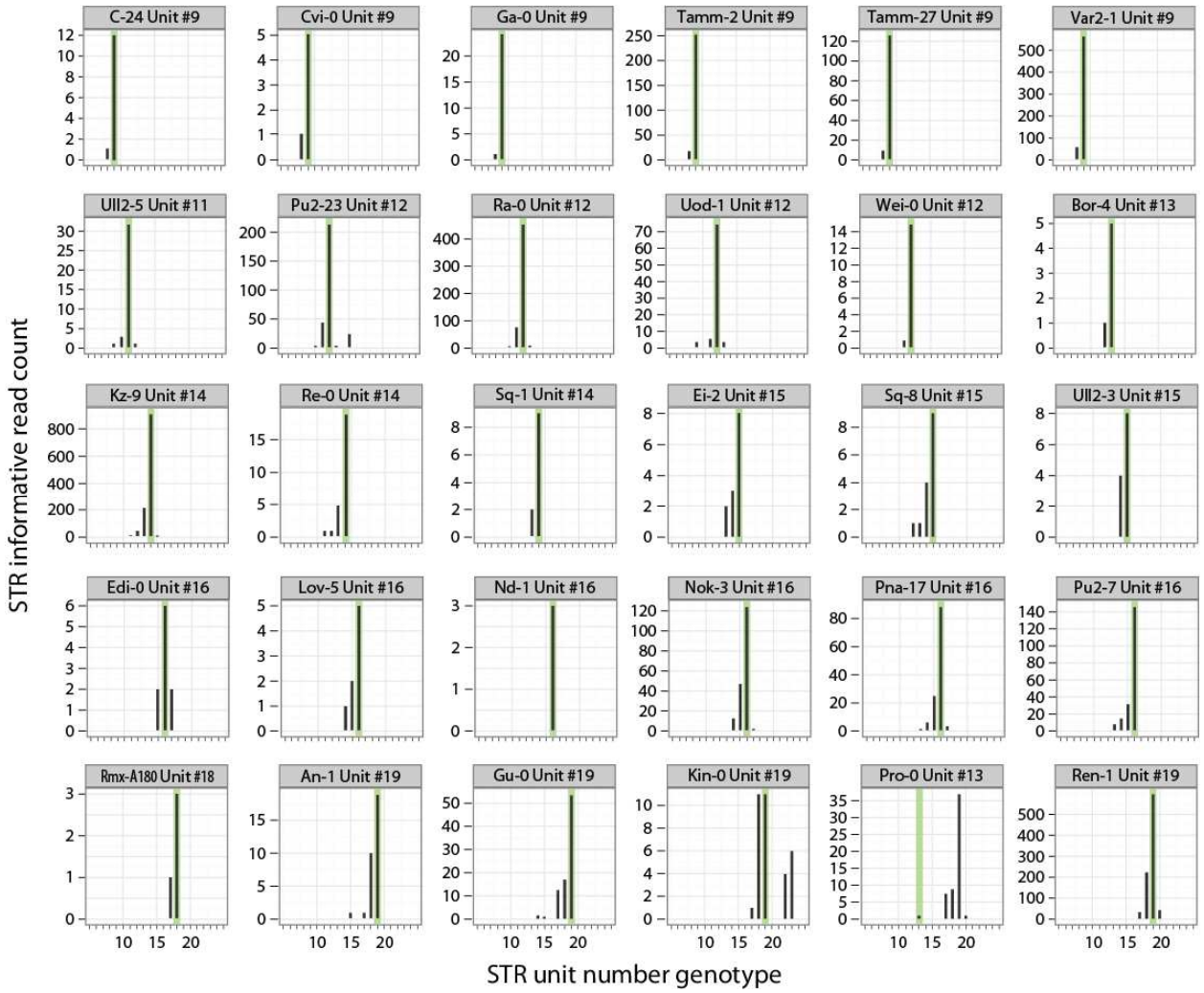

Figure 3.4. MIPSTR accurately determined germ-line *ELF3*-STR unit number on a population-scale across genetically diverse A. thaliana strains. Histograms of raw read counts across 30 accessions. STR unit number as determined by Sanger sequencing is indicated in

green. Using tag-defined read groups, the Kin-0 *ELF3* STR genotype can be resolved to the known STR genotype even with comparatively few total reads. MIPSTR clearly calls STR unit number 19 for Pro-0. Note that different individuals of the same strain were analyzed with MIPSTR and Sanger-sequencing, which may explain the discrepancy.

In general, VARscore correlated well with observed variation across STRs (r=0.68, **Fig. 3.5**), a substantially better agreement than previously observed (Duitama et al. 2014). However, this correlation was substantially weaker among coding STRs (r=0.46) than among non-coding STRs (r=0.75). This discrepancy suggests that sequence characteristics alone do not suffice to predict whether coding STRs vary on a population-scale. Coding STRs are more likely to be functionally important, and thus are less subject to the "neutral model" of the VARscore prediction.

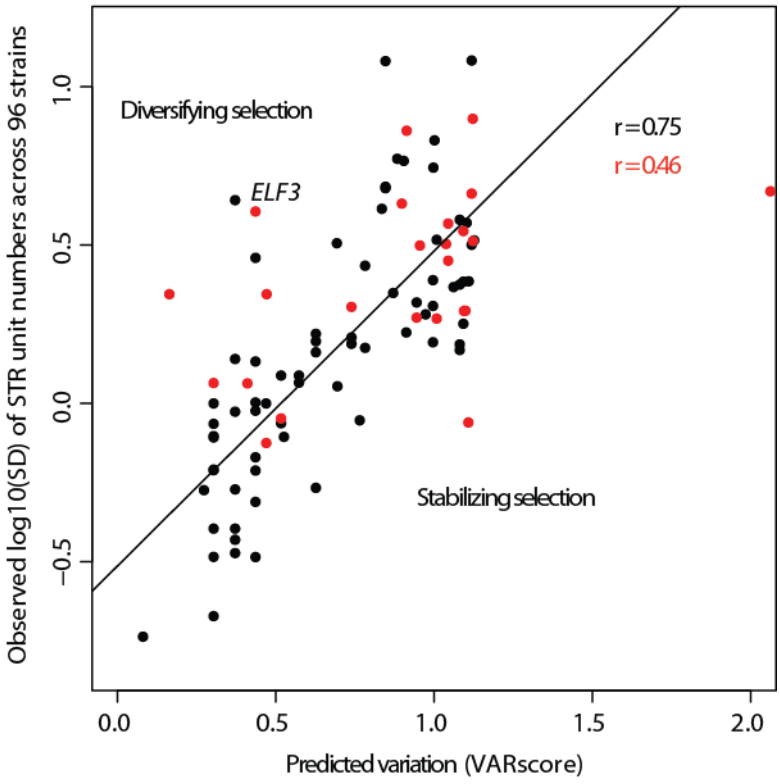

Figure 3.5. Observed and predicted STR variation showed greater correlation for non-coding STRs than coding STRs. The correlation between the observed $\log_{10}$ of the standard deviation of STR unit number across strains (y-axis) and the VARscore (x-axis), which predicts STR variation from sequence characteristics. Black points are non-coding STRs, red points are coding STRs. Outliers may indicate functional importance (*ELF3* STR is indicated).

Deviation of predicted STR variation (*i.e.* VARscore) from observed variation may thus hold information about STR function and selective pressures acting upon it. Specifically, STRs that are observed to be more variable than predicted may be under diversifying selection whereas those STRs that are observed to be less variable than predicted may be functionally constrained and under purifying selection (Press et al. 2014). For example, the STR in the gene *ELF3* is highly variable across strains, ranging from 7 units to as many as 29 units in a set of strains previously analyzed by Sanger sequencing (Undurraga et al. 2012). The phenotypes associated with variation in the *ELF3* STR change dramatically in different genetic backgrounds, suggesting co-evolution of the *ELF3*-STR with epistatically interacting loci (Undurraga et al. 2012a). Given this STR's strong background-dependent phenotypes, it is likely under diversifying selection and, correspondingly, it is much more variable than predicted (**Fig. 3.5**).

A complementary approach for identifying STRs with important function in modulating phenotype is genome-wide association of STR genotypes with phenotypes. The standard statistical methods for associating genotype with phenotype were developed for common, biallelic SNVs (Hayes 2013). STRs are typically multiallelic and often involved in epistatic interactions, both of which make it difficult to associate STR genotype with phenotype using standard methods (Press et al. 2014) . Nevertheless, we performed a naïve association analysis to determine whether STR variation across strains was associated with well-characterized phenotypes (Atwell et al. 2010). We used the one-way analysis of variance (ANOVA) to detect associations between STR loci and phenotypes following previous studies (Mackay et al. 2012), modeling STR alleles as factors to avoid assumptions of linearity (Press et al. 2014). To minimize spurious associations, we dropped STRs that were typed in fewer than 10 strains from this analysis, and for each STR we dropped all strains carrying alleles present in fewer than three strains (rare alleles). We identified 124 significant associations involving 27 STRs and 41 phenotypes at a 1% false discovery rate (**Supplemental Table 3.3**). However, an important caveat is that this analysis did not consider population structure, which is another challenge given the different evolutionary trajectories of SNVs and STRs (Willems et al. 2014).

Our MIP-based approach can easily be scaled to thousands of targets; the human exome MIP set targets ~55,000 loci (Turner et al. 2009). Over 2000 STR loci are accessible by MIPSTR in *A. thaliana*, and many more accessible STR loci exist in humans (Duitama et al. 2014; Guilmatre et al. 2013; Willems et al. 2014; Guilmatre et al. 2013; Molla et al. 2009). Our

preliminary results, considering only a fraction of the accessible *A. thaliana* STR loci, highlight the promise of STRs to contribute to the variation and heritability of quantitative traits (Press et al. 2014).

MIPSTR has potential to sensitively detect heterozygous and somatic STR unit number alleles

To determine the sensitivity with which MIPSTR detects heterozygous and somatic alleles, we mixed DNA of two divergent *A. thaliana* strains, Col-0 and Landsberg (Ler), in known ratios before MIPSTR capture and sequencing (**Fig. 3.6**). Of the 100 STR loci, 56 differed in STR unit number genotypes between Col-0 and Ler, and hence their relative presence across mixtures could be detected by MIPSTR. To assess the relative proportions of STR alleles within each mixture, we determined the number of tag-defined read groups for which the majority of reads supported either the Col-0-specific STR unit number or the Ler-specific STR unit number. This measure, however, is confounded by unequal coverage between libraries. More deeply sequenced libraries will represent a higher number of capture events per target and hence be more likely to identify rare STR alleles (*i.e*. somatic events). To account for variation in number of supporting tag-defined read groups per locus, we performed bootstrap resampling of the modes of the tag-defined read groups at each locus in each library 1000 times, while measuring the proportion of bootstrap samples in which the Col-0 allele was detected. Applying this method to our mixing experiment, the agreement between predicted and observed probabilities of observing Col-0 STR alleles was striking. For example, when we mimicked a "heterozygous" state with a 1:1 Col-0/Ler mixture we observed the Col-0 allele nearly 100% of the time. This agreement of predicted and observed probabilities held across all mixtures (**Fig. 3.6**), indicating that MIPSTR sensitively detects rare alleles. Mixing 1 part Col-DNA into 999 parts Ler-DNA, we were able to detect the Col-alleles at half of the 56 loci.

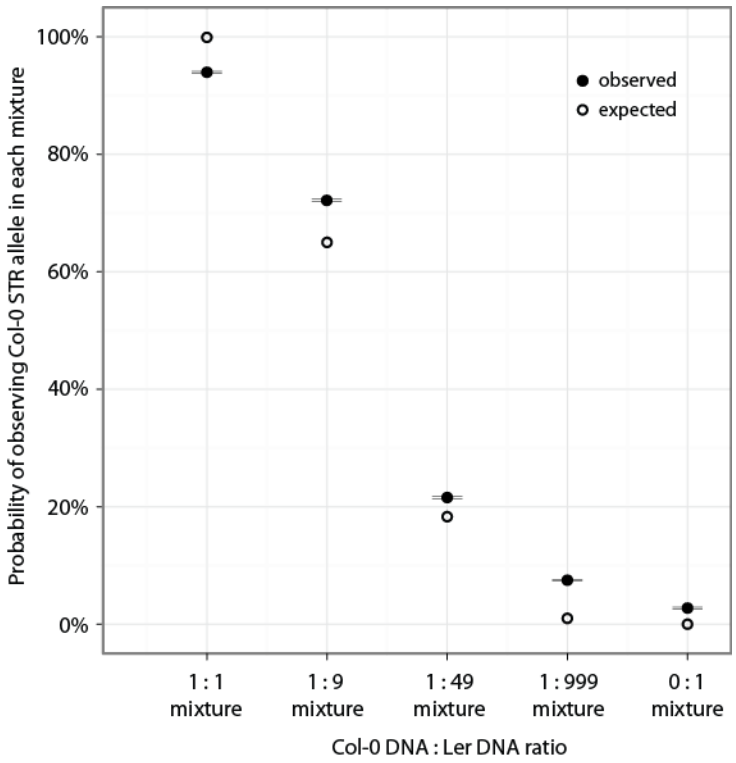

Figure 3.6. MIPSTR detects low frequency STR alleles. X-axis, tested mixtures of Ler and Col-0 DNA. Y-axis, probability of detecting Col-0 STR alleles. Closed circles are observed probability of observing Col-0 STR alleles (standard error is indicated, black lines); open circles are predicted probability of observing Col-0 STR alleles. To calculate the observed probability for each mixture, we re-sampled tag-defined read group modes supporting either the Col-0 or Ler allele at each STR locus 1000 times. The proportion of samples that carry the Col-0 allele was determined and averaged across all STR loci that differ between Ler and Col-0. To calculate the expected probability for each mixture, we assumed the known ratios of Col-0 and Ler STR alleles in each mixture and the probability of observing the Col-0 STR allele with ten observations.

STR instability at selected loci has been previously used as a measure of genome instability and is a hallmark of certain cancers (Kim et al. 2013a; Boland et al. 1998a). Our data suggest that MIPSTR has the potential to offer considerably greater resolution by assessing somatic STR variation genome-wide. To examine the potential of our method to detect decreased genome stability, we performed MIPSTR on *Atmsh2* mutant plants. This mutant carries an insertion in the *MSH2* gene, which is a crucial component of the DNA repair machinery. Indeed, a previous study, using a reporter system, found a ~10% increase in dinucleotide STR somatic mutation events in this mutant (Golubov et al. 2010). We applied MIPSTR to three Col-0 plants

and three *Atmsh2* plants. After eliminating STR loci with high technical error rates and loci without information for both strains, we compared the average number of STR alleles per locus with bootstrap resampling as described above. Instead of assessing two alleles, those of Col-0 and Ler as in the mixtures, we counted all alleles supported by at least one tag-defined read group in the resampling procedure. Compared to Col-0, the *Atmsh2* plants showed a 4.7% increase in average STR alleles across loci ($p < 2.2E-16$, Wilcoxon test, **Supplemental Fig. 3.4A**). Removing the two most overrepresented Col-0 and *Atmsh2* libraries, (*i.e.* with many more tag-defined read groups represented), resulted in an even larger difference between Col-0 and *Atmsh2*, with a 10.6% increase in *Atmsh2* mutants' average STR alleles across all tested loci ($p < 2.2E-16$, Wilcoxon test, **Supplemental Fig. 3.4B**). This result is particularly remarkable considering that these loci were not optimized with respect to those most likely to exhibit somatic variation. Such optimization is readily possible with MIPSTR – for example, by applying MIPSTR to long non-coding dinucleotide STRs, which are far more prone to unit number mutation and hence somatic error. By combining such a specifically designed set of smMIPs (*i.e.* targets) for detecting somatic STR variation with deep sequencing, MIPSTR may be capable of identifying much more subtle increases in genome instability.

Discussion

The potential of STR variation to contribute to phenotypic variation and heritability of complex traits is increasingly recognized (Press et al. 2014). To realize this potential, several recent efforts, relying on either analytical or experimental innovation, have made progress towards the ascertainment of accurate STR genotypes on a population-scale (Cao et al. 2014; Duitama et al. 2014; Guilmatre et al. 2013; Gymrek et al. 2012; Highnam et al. 2013). However, the STR-specific challenges for accurate genotyping – mappability and high amplification stutter – were only partially addressed. Here, we resolve these challenges by capturing STRs with single-molecule Molecular Inversion Probes that allow detection of many independent capture events of the same STR across many DNA molecules (Hiatt et al. 2013). Specifically, we resolve the mappability challenge by using targeted capture and locus-specific synthetic reference sequences. We resolve the challenge of inherently high technical error in STR amplification by examining many tag-derived read groups for each STR locus. STR unit number variation within a tag-defined read group results from amplification stutter. In contrast, STR unit number

variation among tag-defined read groups has the potential to detect genomic duplications, heterozygosity, and somatic variation. We show that MIPSTR is capable of distinguishing these crucial sources of STR variation within samples.

Previous studies relied on amplification of haploid or homozygous genomes to estimate technical error for STR-containing sequencing reads (Guilmatre et al. 2013; Gymrek et al. 2012; Highnam et al. 2013); this approach is confounded by somatic variation and high STR mutation rates. MIPSTR offers an experimental avenue for empirically ascertaining technical error for many types of STRs. Notably, we observed dramatic differences in technical error even among the 100 trinucleotide and hexanucleotide STRs tested here. With larger numbers and more types of STRs, one may derive more precise predictions of sequencing error based on sequence composition, length, genomic position and other features.

However, even in this proof-of-principle study some patterns emerged that inform our understanding of the mutability of STRs. First, as others have seen, the most common technical error we observed was the loss of one STR unit (STR variation within tag-defined read groups) (Guilmatre et al. 2013). The loss of one STR unit was also the most common somatic event (STR variation observed among tag-defined read groups). As STR variation within a tag-defined read group exclusively derives from amplification stutter, we speculate that the somatic loss of one STR unit similarly derives from amplification errors during replication, rather than errors in DNA recombination or repair. Second, as anticipated by previous studies (Legendre et al. 2007), longer STRs showed both increased technical and somatic error. Third, comparing predicted (based on neutral models) to observed variation in STR unit number, we found a stronger correlation for non-coding STRs than coding STRs, consistent with greater selective pressures on the latter, and suggesting that deviations from expected STR variation may hold information about an STR's functional importance.

Although the immediate application of MIPSTR is in accurately assessing germ-line STR variation, we also emphasize our method's potential to sensitively detect somatic STR variation. Somatic STR variation, better known as microsatellite instability (MSI), has a long history as a biomarker for certain colorectal cancers, more recently also for endometrial cancers (Boland et al. 1998; Kim et al. 2013). In fact, a recent study used exome sequencing data (~20X coverage, 100 bp reads, compare with Figure S 3.1) to assess MSI in colorectal and endometrial tumor and matched normal samples (Kim et al. 20133). Using only STR-spanning reads, this study called

an MSI event at a given STR locus by comparing STR unit number distributions between tumor and matched normal samples, controlling for technical error with the STR variation observed in normal samples. As we show, comparing read distributions is vulnerable to differences in coverage and requires normalization by bootstrap resampling. MIPSTR eliminates the need to compare distributions of “normal” and ‘tumor’ samples to correct for technical error because MIPSTR calls both germ-line STR genotype and somatic STR variation in a given sample.

Although the STR loci that we targeted were not optimized for somatic events, MIPSTR detected the Col-0 STR alleles even in a 1:999 mixture of Col-0 and Ler-DNA. Moreover, using MIPSTR we observed a substantial increase of somatic events in a plant mutant deficient in DNA repair. MIPSTR can readily test and identify panels of 100-500 STR loci that are particularly unstable and prone to many somatic mutation events – for example by testing longer and less complex STRs such as di- or mononucleotides.

Beyond cancer genomics, at a population-scale, somatic variation and its occurrence across tissues, developmental stages, and in response to environmental perturbations has remained largely inaccessible due to the prohibitive costs of ultra-deep and single-cell sequencing (Baslan et al. 2012; Navin et al. 2011). As STRs are highly mutable, they are arguably the best biomarkers to detect even subtle perturbations of genome stability. We suggest that MIPSTR in combination with STR panels optimized for somatic variation has great promise to detect even subtle decreases in genome stability. We and others have previously proposed that subtly decreased genome stability may precede or coincide with many disease processes and may increase the penetrance of disease risk alleles (Queitsch et al. 2012; Heng 2010; Poduri et al. 2013). MIPSTR offers an approach to empirically test this hypothesis. Compared to single cell sequencing (Baslan et al. 2012; Navin et al. 2011). MIPSTR also offers a cost-and labor-efficient alternative for assessing the genetic heterogeneity of tumors, which is clinically relevant for disease treatment and prognosis (Fox et al. 2013; Schmitt et al. 2012)

Finally, we emphasize that MIPSTR is readily scalable: by simply targeting all STR loci in its size range, our method can provide genome-wide assessment of STR variation; by sequencing more deeply for an optimized panel of STR loci our method can provide information about somatic variation. MIPSTR is applicable to any organism with a high-quality reference genome, including humans. In the future, applying MIPSTR across populations of diverse

species will contribute to fulfilling the long overdue promise of STR variation for explaining trait heritability.

Methods

smMIP capture reagent design

Each smMIP is an 80 bp oligonucleotide with a 40 bp common backbone flanked by an extension arm of 16-20 bp and a ligation arm of 20-24 bp. These unique arms specifically hybridize to flanking regions of STR loci for a gap-fill of 200 bp. Included in the 40 bp of the common backbone are 12 random nucleotides, the degenerate tag, generating ~12^4 = 1.67 x 10^6 unique sequences per MIP. The MIPs were designed for 102 STRs across the *A. thaliana* genome (**Supplemental Table 3.1**).

These MIPs were procured individually by column-synthesis on an 100 nmol scale with standard desalting purification (at a cost of ~$32 per MIP). Once purchased, one has effectively an infinite MIP supply allowing for millions of capture reactions, justifying the considerable upfront MIP cost. Cost per MIP is significantly lower when ordering less MIP without purification (25 nmol /$7.20 per MIP) (Hiatt et al. 2013).

MIPs were pooled at equal molarity and mixed with the target at 200-fold molar excess. The results of the first capture reaction in the Col-0 reference genome, specifically the distribution of read counts from each MIP, were used to adjust MIP concentrations. We increased the concentration of the lowest performing MIPs (28, fewest number of reads) 100-fold; concentration of the next lowest performing group of MIPS (43) was increased 10-fold.

Capture and Library Construction

DNA was extracted from rosette leaves of individual 20-day-old *A. thaliana* plants using DNeasy Plant Maxi Kit (Qiagen). DNA was cleaned up and concentrated with Amicon Ultra Centrifugal Filter Units (Millipore).

Capture procedures were modified from previous protocols (O'Roak et al. 2012; Hiatt et al. 2013). 750 ng genomic DNA was mixed with 2 pmol smMIP mixture (starting concentration before adjustment for low performing MIPs), 1.5 µl 10X Ampligase buffer, and molecular biology grade water to a total volume of 15 µl. For hybridization, these mixtures were incubated in a thermocycler with a heated lid for 10 minutes at 95°C followed by 48 hours at 55°C. After hybridization, we added 2.5 pmol dNTPs (TaKaRa), 1 unit Ex Taq polymerase (TaKaRa), 0.5 µl

10x Ampligase buffer, 60 units Ampligase DNA ligase (Epicentre) and molecular grade water to an added volume of 5 μl per mixture. The extension phase was carried out at 60°C for an hour. After gap-fill and ligation, the mixtures were cooled to 37°C for two minutes. We then added 40 units of Exonuclease I (NEB) and 200 units of Exonuclease III (NEB) for a total reaction volume of 19 μl. To digest uncircularized and excess genomic DNA, we incubated these mixtures at 37°C for 15 minutes, and then denatured the enzymes at 92°C for two minutes.

Library construction, purification, and pooling

To create sequencing libraries, we amplified the capture reactions using a common forward primer and an indexed reverse primer. We mixed 5 μl capture reaction with 12.5 pmol dNTPs (TaKaRa), 5 ul 10X Ex Taq buffer, 25 micromoles forward primer, 25 micromoles reverse primer, 1 unit Ex Taq polymerase (TaKaRa), and molecular biology grade water to a total reaction volume of 50 μl. We performed an initial denaturation at 98°C for 10 seconds, followed by 28 cycles of 10 seconds at 98°C, 30 seconds at 58°C and 12 seconds at 72°C. The final extension was for 3 minutes at 72°C. PCR products were pooled as equal volumes per sample or according to gel image quantification to get approximately equal representation. We then cleaned up the pooled PCR products using Ampure XP beads (Agencourt) at 1.8X according to manufacturer's recommendations.

Sequencing and primary analysis

Samples were sequenced using the Illumina MiSeq v2 platform according to the manufacturer's instructions with custom sequencing primers (Hiatt et al. 2013). To improve cluster generation for these low complexity STR libraries, we spiked in Phi-X or whole genomic DNA libraries at 10-20%. We collected one 250 bp forward read to determine sequence of the ligation arm and STR target locus, one 50 bp reverse read to determine the sequence of the degenerate tag and extension arm, and one 8 bp read to determine the sample index sequence. The MiSeq software sorted by index read to separate pooled libraries.

Mapping and STR genotype calling

For each target STR locus, we created a synthetic reference of 100 "chromosomes," which consisted of the Col-0 reference target sequence with 1 to 100 pure STR units (no SNVs). We sorted reads by the first 16 bp of the ligation targeting arm allowing three mismatches and then used the bwasw alignment mode of the *bwa* aligner (Li and Durbin 2009) to map the reads to the locus-specific synthetic reference. For a given read, if the A-score of its alignment to a

specific synthetic "chromosome" was ≥180, we called the STR unit number of this "chromosome" for this read. Below this A-score, the read was discarded. When the sequence read ended within the STR (presumably due to a large expansion of STR units) but still mapped with an acceptable A-score, we called the genotype as ≥ the unit number of the "chromosome" to which the read aligned. In this way, MIPSTR can yield information about STR unit number expansions in a given individual even in the absence of STR-spanning reads. Here, these "≥" calls were not used in further analyses such as association or calculation of variation.

We then sorted the STR genotype calls by the degenerate tag on the paired reverse read from which they derived. We required an exact match of the 12 bp degenerate tag for reads to be grouped into a tag-defined read group. We then called the mode STR unit number of each tag-defined read group as the genotype of that DNA molecule. If we observed that more than one tag-defined read group supported an alternate STR allele, we considered it evidence of somatic variation.

STR association with phenotypes

We used previously published data for 107 phenotypes collected for 96 *A. thaliana* strains (Atwell et al. 2010). We then proceeded to detect associations between each of these phenotypes and each variable STR locus within genotyped strains. For each test, we omitted strains from the analysis that were not phenotyped for the relevant trait or genotyped at the STR in question. We additionally removed from the analysis strains that carried STR alleles that were found in fewer than three strains total, to avoid confounding from rare alleles. We then performed one-way ANOVA to test the null hypothesis of no association between each STR and each phenotype, while treating each STR allele categorically. We chose to treat STR alleles categorically because assumptions of linearity in STR-phenotype associations are poorly founded in some cases (Undurraga et al. 2012; Press et al. 2014). Associations were accepted at a 1% false discovery rate ($p = 1.48 * 10^{-4}$).

Calculating technical error rates

To calculate the technical error rate of amplifying STR loci, we considered all tag-defined read groups for which a single STR unit number mode was supported by at least two reads. For these tag-defined read groups, we took the fraction of reads supporting unit numbers other than the mode and divided by the total number of reads. We averaged across all tag-defined

read groups at a given locus for a technical error score between 0 and 1 representing the fraction of reads at a locus known to be error (**Supplemental Table 3.1**).

Somatic allele counts

To compare the number of somatic events occurring in different individuals, we only considered STR loci with low technical error scores (below 0.2, **Supplemental Table 3.1**) and with information for all plants in the comparison. We used bootstrap resampling to account for sometimes vastly different read counts. For example, in the Col-0 and Ler mixing experiment, some mixture libraries had as few as ten tag-defined read groups at a given locus. Thus, we resampled ten modes from tag-defined read groups in these samples, counting the proportion of those samples in which the Col-0 unit number allele was present. In Col-0 versus *atmsh2* experiment, depth of coverage was much higher and hence we resampled 1000 modes of tag-defined read groups for each locus. For each sample, we calculated how many different STR unit number alleles were present and averaged across loci.

Acknowledgements

This work was supported by grants from the National Human Genome Research Institute Interdisciplinary Training in Genomic Sciences (2T32HG35-16 to MOP, T32 HG00035 to KDC) and the National Institute of Health New Innovator Award (DP2OD008371 to CQ). We would like to thank Matt Rich, Josh Cuperus, Matthew Snyder, Akash Kumar, Joe Hiatt, Choli Lee, Rachel Youngblood, Giang Ong, Jacob Kitzman, and Queitsch lab members for helpful discussions.

Supplemental Figures

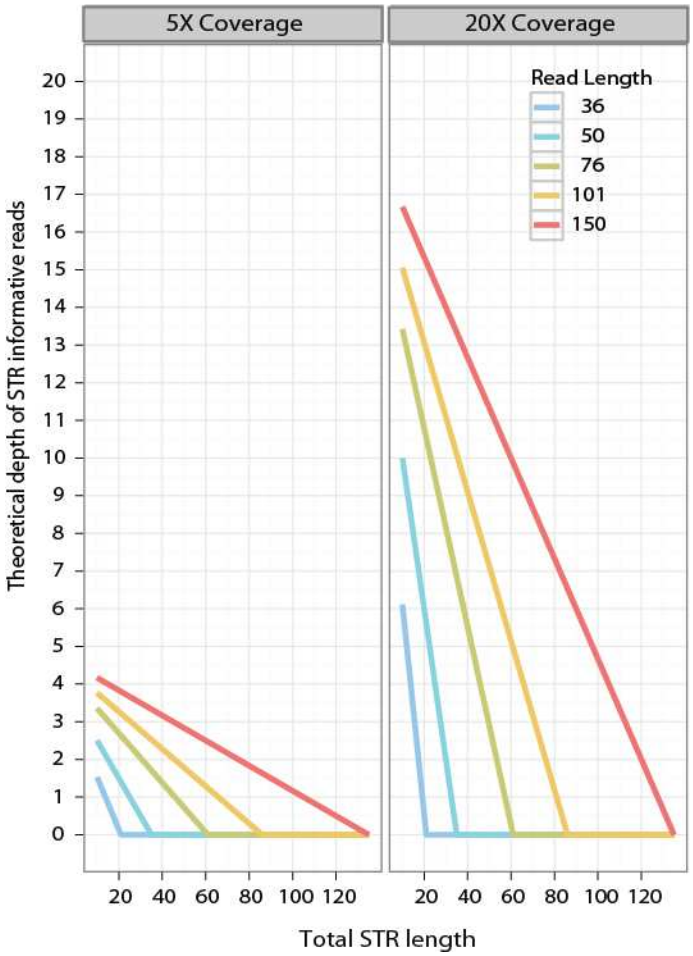

Figure S3.1. Theoretical number of STR informative reads given different read lengths, sequencing coverage, and total STR length. For example, at 20X sequencing coverage with 101bp read length, a 60bp STR will be represented by only six STR-spanning, informative reads.

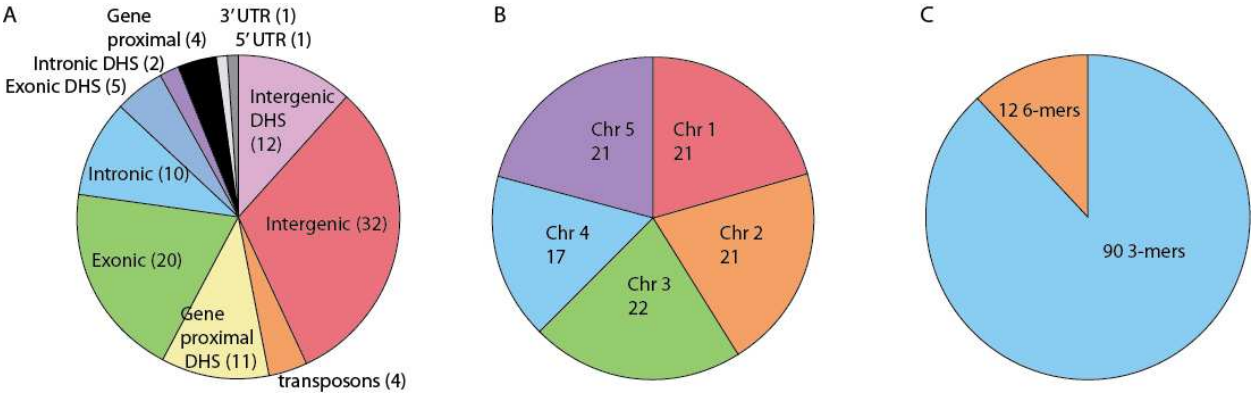

Figure S3.2. Characteristics of smMIP target STRs. A) Distribution in genomic functional regions. The four STRs located in transposons consist of two pairs of duplicated loci, both members of a pair are targeted with the same probe. Gene proximal STRs are within 1 kb of a gene. B) The targets are distributed evenly throughout the genome. C) The targets are all tri- or hexa- nucleotide STRs. The duplicated STR loci are hexamers.

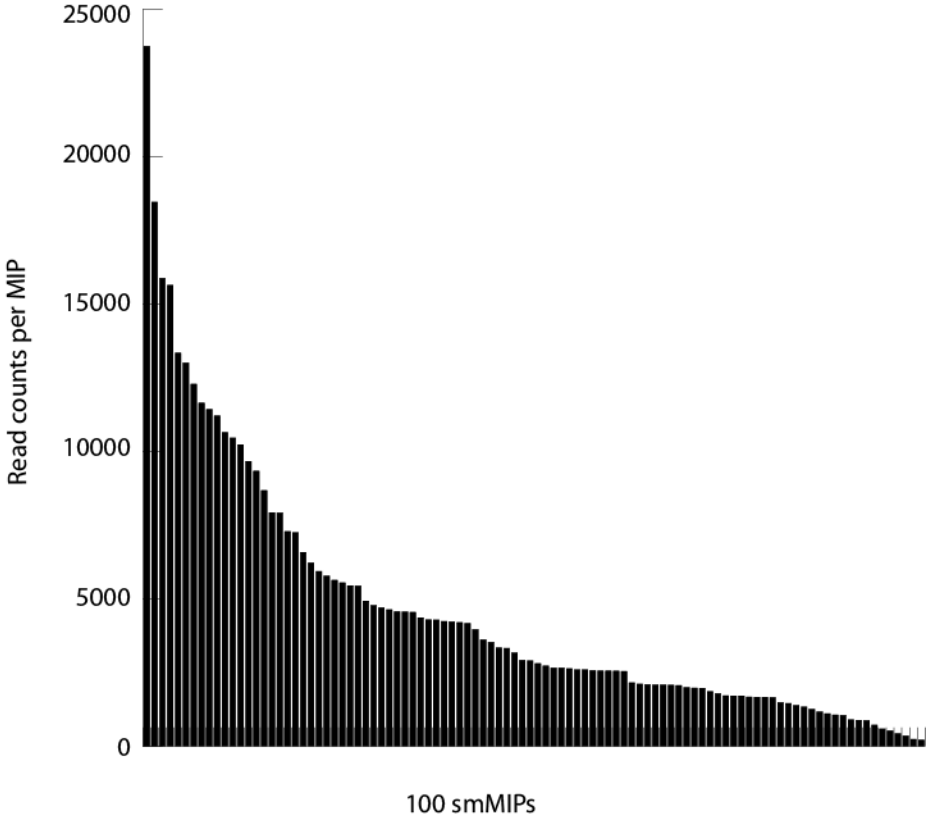

Figure S3.3. MIPSTR successfully captured all 102 STR target loci. X-axis, 100 smMIPs targeting different STRs. Y-axis, read count for each smMIP. As previously observed, smMIP capture efficiency differed among probes (Hiatt et al. 2013).. Note that two smMIPs capture duplicate loci (100 smMIPs, 102 STR target loci).

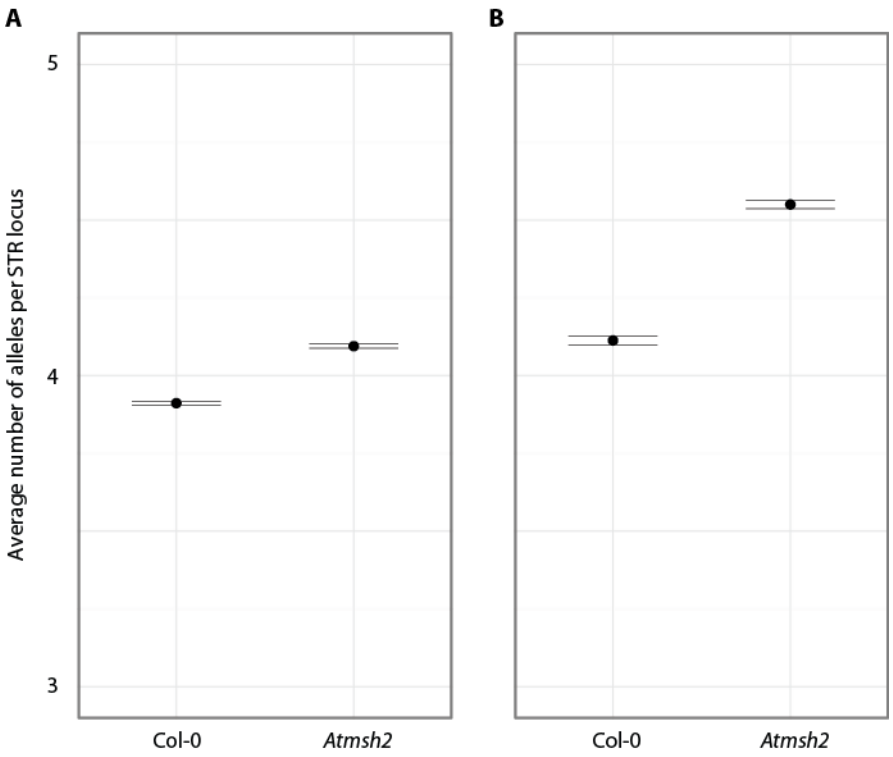

Figure S3.4. MIPSTR detects decreased genome stability in *Atmsh2* mutants. X-axis, Col-0 control or *Atmsh2* mutant plants. Y-axis, average number of alleles per STR locus (no somatic variation in homozygous plants equals one allele). Standard error is indicated with black lines. To calculate the average number of alleles per locus, we re-sampled tag-defined read group modes at each STR locus 1000 times. The number of alleles supported by at least one tag-defined read group per sample was averaged across STR loci. **A**) Data for three Col-0 plants and three *Atmsh2* plants. **B**) Data for two Col-0 plants and two Atmsh2 plants with the outliers removed (greater than two-fold higher depth of coverage).

CHAPTER FOUR

LESSONS FROM MODEL ORGANISMS: PHENOTYPIC ROBUSTNESS AND MISSING HERITABILITY IN COMPLEX DISEASE[3]

Abstract

Genetically tractable model organisms from phages to mice have taught us invaluable lessons about fundamental biological processes and disease-causing mutations. Owing to technological and computational advances, human biology and the causes of human diseases have become accessible as never before. Progress in identifying genetic determinants for human diseases has been most remarkable for Mendelian traits. In contrast, identifying genetic determinants for complex diseases such as diabetes, cancer, and cardiovascular and neurological diseases has remained challenging, despite the fact that these diseases cluster in families. Hundreds of variants associated with complex diseases have been found in genome-wide association studies (GWAS), yet most of these variants explain only a modest amount of the observed heritability, a phenomenon known as "missing heritability." The missing heritability has been attributed to many factors, mainly inadequacies in genotyping and phenotyping. We argue that lessons learned about complex traits in model organisms offer an alternative explanation for missing heritability in humans. In diverse model organisms, phenotypic robustness differs among individuals, and those with decreased robustness show increased penetrance of mutations and express previously cryptic genetic variation. We propose that phenotypic robustness also differs among humans, and that individuals with lower robustness will be more responsive to genetic and environmental perturbations and hence susceptible to disease. Phenotypic robustness is a quantitative trait that can be accurately measured in model organisms, but not as yet in humans. We propose feasible approaches to measure robustness in large human populations, proof-of-principle experiments for robustness markers in model organisms, and a new GWAS design that takes differences in robustness into account.

---

[3] This chapter is published as "Lessons from model organism: phenotypic robustness and missing heritability in complex disease," *PLoS Genet*: 8(11), 2012, by C Queitsch, KD Carlson, S Girirajan.

Introduction

Complex diseases such as diabetes, cancer, and cardiovascular and neurological diseases are the predominant causes of morbidity and mortality in the developed world. As they tend to cluster in families, these diseases are thought to involve genetic factors in addition to environmental ones. Although genome-wide association studies (GWAS) have identified hundreds of common variants associated with complex diseases and although susceptibility loci have been reported for many disorders, the overall genetic risk explained by these loci remains modest. Thus, only a small proportion of heritability (proportion of phenotypic variance explained by genetic variants) has been accounted for (Eichler et al. 2010; Manolio et al. 2009). This discrepancy, termed 'missing heritability,' has been attributed to many factors. First, rare variants of large effect size (odds ratio > 2) may account for some of the unexplained genetic risk, as observed in neuropsychiatric disorders such as autism, schizophrenia, and developmental delay (Eichler et al. 2010; Manolio et al. 2009; Girirajan et al. 2011b; Girirajan and Eichler 2010). By their nature, rare variants are difficult to detect and to associate with phenotype using statistics. Second, highly repetitive structural and sequence variants have remained inaccessible to large-scale genotyping (Eichler et al. 2010; Manolio et al. 2009; Gibson 2011). Third, heritability estimates may be artificially inflated due to interactions between genes, to shared environments in families, and to imprecise diagnoses of complex disorders (Eichler et al. 2010; Manolio et al. 2009; Gibson 2011; Zuk et al. 2012). Consequently, current research addresses the problem of missing heritability with more comprehensive genotyping of genetic variants in statistically well-powered cohorts of individuals that are better characterized for disease phenotypes, genetic background, and environmental exposure (Eichler et al. 2010; Manolio et al. 2009). This approach is rooted in the prevalent hypothesis that some combination of rare variants of large effect, common variants of small effect, and environmental factors will translate into disease (Eichler et al. 2010; Manolio et al. 2009; Gibson 2011).

An alternative view posits that cryptic genetic variation accounts for a substantial fraction of disease-associated risk (Gibson 2009). In robust individuals, cryptic genetic variation will not contribute to disease and will elude detection by GWAS. In contrast, in individuals with decreased overall phenotypic robustness, formerly cryptic genetic variants will contribute to disease, and disease-related variants will increase in penetrance, resulting in increased heritability. This hypothesis draws on findings from diverse model organisms including yeast,

worms, flies, plants, and fish: decreased phenotypic robustness significantly increases heritability of complex traits due to revealed, formerly cryptic genetic variation and increased penetrance of genetic variants (Queitsch et al. 2002; Rutherford et al. 2007; Rutherford and Lindquist 1998; Sangster et al. 2004, 2008, 2008; Yeyati et al. 2007; Burga et al. 2011; Casanueva et al. 2012). In this review, we describe the causes and consequences of decreased phenotypic robustness in model organisms, relate these findings to complex disease phenotypes, and propose an alternative GWAS approach that accounts for differences in robustness among humans.

What is phenotypic robustness?

Phenotypic robustness is the ability of a given genotype to produce a constant phenotype, even when the organism is faced with genetic or environmental perturbations. The remarkable phenotypic robustness of wild-type organisms is commonly attributed to features of the underlying genetic networks, such as modularity, feedback loops, gene redundancy, connectivity, degeneracy, and the presence of activity-modulating microRNAs (Rutherford et al. 2007; Sangster et al. 2004; Bergman and Siegal 2003; Ciliberti et al. 2007; de Visser et al. 2003; Hornstein and Shomron 2006; Jarosz et al. 2010; Leclerc 2008; Lehner 2008; Levy and Siegal 2008; Li et al. 2009; Manu et al. 2009; Masel and Siegal 2009; Raser and O'Shea 2005; Wagner 2000, 2008; Salathia and Queitsch 2007; Baggs et al. 2009; Whitacre 2012). In model organisms, targeted perturbation of any of these features can decrease phenotypic robustness and release cryptic genetic variation (Queitsch et al. 2002; Rutherford and Lindquist 1998; Yeyati et al. 2007; Burga et al. 2011; Bergman and Siegal 2003; Levy and Siegal 2008; Li et al. 2009; Manu et al. 2009; Baggs et al. 2009; Jarosz and Lindquist 2010; Lehner et al. 2006).

Phenotypic robustness is a measurable quantitative trait. Traditionally, robustness of individuals has been measured as the degree of symmetry in morphological features (Debat and David 2001). A high degree of symmetry is thought to be associated with high fitness and even with the perception of beauty for human faces (Gangestad et al. 2005; Scheib et al. 1999; Thornhill and Gangestad 1999). In most organisms, objective and high-throughput analysis of symmetry is complicated by the complexity of morphological features and their profound changes throughout development. Another measure of robustness is the degree of accuracy with which a genotype produces a particular phenotype across many isogenic siblings (Debat and David 2001). By this measure, phenotypic robustness, like any other quantitative trait, shows a

distribution among genetically divergent individuals of a species and can be mapped to distinct genetic loci (Sangster et al. 2008). The ability to buffer mutations can vary among isogenic individuals, suggesting that non-genetic mechanisms significantly affect robustness (Burga et al. 2011; Casanueva et al. 2012). These non-genetic robustness determinants will elude genetic approaches. None of the robustness measures that have been used in model organisms are applicable in humans; however, they have proven useful to identify master regulators or network hubs that may contribute to robustness in humans.

The best-characterized master regulator of robustness is the molecular chaperone HSP90 (Queitsch et al. 2002; Rutherford et al. 2007; Rutherford and Lindquist 1998; Sangster et al. 2008, 2008; Yeyati et al. 2007; Burga et al. 2011; Jarosz and Lindquist 2010; Jarosz et al. 2010; Gangaraju et al. 2011; Sollars et al. 2003; Specchia et al. 2010b; Taipale et al. 2010). HSP90 assists the proper folding and function of many key enzymes and transcription factors that govern growth and development (Taipale et al. 2010). The chaperone is essential in eukaryotes, evolutionarily conserved, highly connected, and plays a crucial role in integrating environmental signals (Taipale et al. 2010). HSP90's function is of even greater importance under environmental stress that compromises protein folding (Taipale et al. 2010). In genetically divergent plant, fly, yeast, and fish populations, HSP90 inhibition significantly increases heritability due to increased penetrance of known genetic variants and revealed cryptic genetic variants (Queitsch et al. 2002; Rutherford and Lindquist 1998; Yeyati et al. 2007; Jarosz and Lindquist 2010). In worms, naturally varying HSP90 levels predict mutation penetrance: lower HSP90 levels result in greater penetrance of mutations (Burga et al. 2011; Casanueva et al. 2012). In plants, yeast, and flies, HSP90-dependent variants are common in natural strains and often affect complex traits (Sangster et al. 2008a, 2008b; Jarosz and Lindquist 2010; Carey et al. 2006).

Decreased phenotypic robustness is associated with genome instability

Loss of robustness may be associated with an increased mutation rate. In flies, HSP90 inhibition increases transposon transcription and mobility (Gangaraju et al. 2011; Specchia et al. 2010). In human cells, HSP90 inhibition compromises repair of DNA damage in response to radiation (Dote et al. 2006) and increases the mutation rate of microsatellites (Mittelman et al. 2010). In yeast, severe HSP90 inhibition induces aneuploidy (Chen et al. 2012). In plants,

HSP90 inhibition increases the rate of homologous recombination (**Fig. 4.1**). HSP90 inhibition appears to interfere broadly with genome stability by affecting transposon silencing, DNA repair, microsatellite stability, chromosome segregation, and homologous recombination. Given the extent of standing variation that responds to HSP90 inhibition (Sangster et al. 2008b, 2008a; Jarosz and Lindquist 2010; Carey et al. 2006), newly arising mutations probably play a minor role in HSP90-dependent phenotypes, yet genome instability may be a hallmark of decreased robustness. These HSP90-specific results are consistent with stress-induced increases in mutation rate in bacteria, yeast, and plants (Chen et al. 2012; Molinier et al. 2006; Boyko and Kovalchuk 2011; Ponder et al. 2005; Shee et al. 2011; Cooley et al. 2010). Environmental stress also decreases robustness in diverse organisms, supporting an association of robustness with genome stability (Queitsch et al. 2002; Rutherford and Lindquist 1998; Yeyati et al. 2007; Dworkin 2005; Parsons 1992).

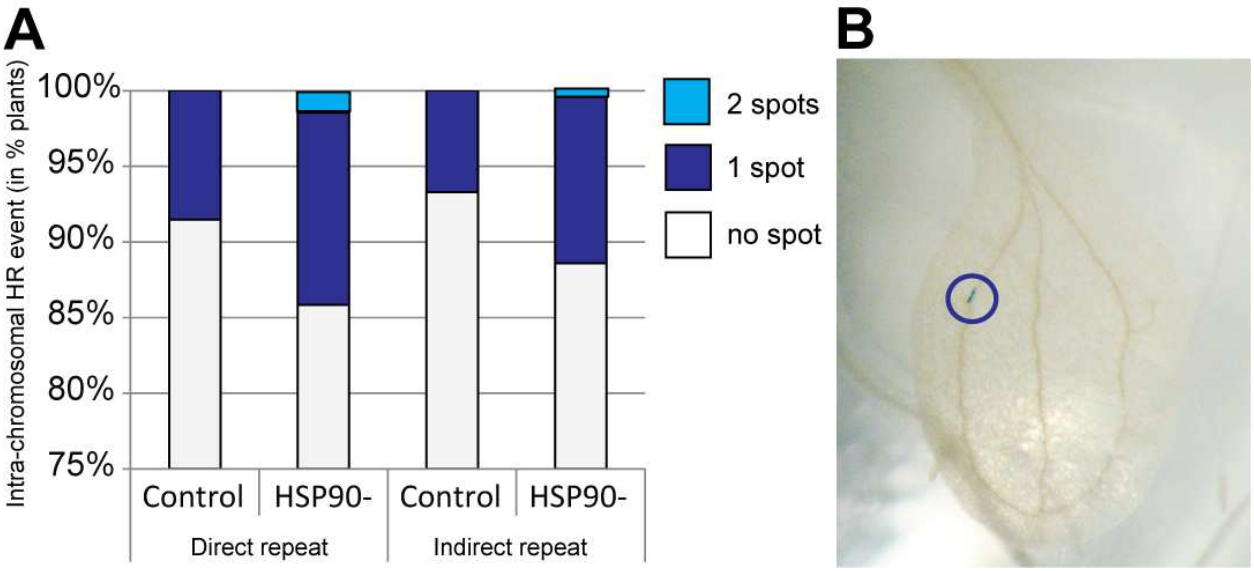

Fig. 4.1: HSP90 inhibition increases homologous recombination in plants. A. Transgenic plants carrying reporter constructs that monitor somatic homologous recombination (HR) events [48] were grown with and without HSP90 inhibition. HR restores a functional GUS gene, allowing for detection of somatic HR events. Intra-chromosomal HR events were significantly increased in plants grown with the HSP90 inhibitor geldanamycin for direct and indirect repeat reporter constructs (Poisson regression p=0.0078). A total of 925 seedlings were analyzed with 381 seedlings for the direct repeat reporter line and 544 for the indirect repeat reporter line. **B.** Example of somatic HR event in plant leaf (blue circle around GUS spot). Unpublished data by K. Carlson, A. Nuttle, and C. Queitsch.

HSP90 is only one of several ‘robustness’ master regulators. In yeast, 60-300 genes, all network hubs, have been identified for which deletion of any one decreases robustness (Levy and Siegal 2008). In worms, decreased function of highly connected hub genes enhances the phenotypic consequences of decreased function in many other genes (Lehner et al. 2006). It remains untested how many of these yeast and worm network hubs reveal cryptic genetic variation in divergent populations or decrease genome stability when perturbed. However, effects on genome instability seem likely as both worm and yeast network hub genes are strongly enriched for chromatin regulators (Levy and Siegal 2008; Lehner et al. 2006). In plants, a novel master regulator of robustness has been found, in which loss of function decreases phenotypic robustness and releases cryptic genetic variation (Cristina Alexandre, personal communication). Consistent with an association of robustness and genome stability, inhibition of this master regulator increases transposon mobility.

Release of cryptic genetic variation may contribute to rise of complex diseases

While both rare variants of large effect and many common variants of small effect contribute to complex diseases (Gibson 2011), neither model satisfactorily explains the significant rise of complex diseases (Gibson 2011, 2009). In the last century, dramatic changes in life style and environment included diet changes, refrigeration, departure from natural circadian rhythm through artificial lighting, modern hygiene, and urban living, to name a few (Gibson 2009; Hugot et al. 2003; Strachan 2000b, 2000a). In particular, changes in diet and refrigeration have led to the fast evolution and changing composition of the human gut microbiome (Greenblum et al. 2012; Ley et al. 2008a, 2008b; Malekzadeh et al. 2009; Prentice et al. 2008; Turnbaugh et al. 2008, 2009). Greg Gibson has suggested that these environmental perturbations may alter the genetic contributions to phenotype by revealing cryptic genetic variation, especially among individuals with reduced phenotypic robustness (Gibson 2011, 2009).

Gibson’s hypothesis that cryptic genetic variation contributes to disease susceptibility is supported by the properties of disease risk variants and their distribution among human populations. A surprising number of the single nucleotide polymorphisms (SNPs) associated with disease risk are ancestral, indicating that the protective variant arose in the human lineage. Thus, disease susceptibility cannot be easily explained by acquisition of deleterious mutations in the human lineage. Further, disease risk alleles vary dramatically in frequency and effect size in

human populations, indicating extensive population structure and the importance of environmental factors for developing disease (Gibson 2009).

Features of certain complex diseases are consistent with decreased robustness

As common SNPs fail to confer significant disease risk for disorders such as autism, schizophrenia, and mental retardation, rare variants of potentially large effect have been examined (Eichler et al. 2010; Manolio et al. 2009). The human genome contains regions that are predisposed to copy number variation (CNV) due to their repeated architecture (Girirajan et al. 2011b). Several of these rare, recurrent CNVs are associated with schizophrenia, autism, cardiac and renal anomalies, epilepsy, obesity, diabetes, and mental retardation (Girirajan et al. 2011b; Girirajan and Eichler 2010; Bailey et al. 2002; Girirajan et al. 2011a, 2010; Sebat et al. 2007; Sharp et al. 2006; Walsh et al. 2008). However, many of these CNVs are also found in control populations and unaffected family members. Moreover, the same CNV can be associated with a large spectrum of disorders (Girirajan et al. 2011b; Girirajan and Eichler 2010). For example, del17q12 is associated with renal cysts, maturity-onset diabetes, developmental delay, brain malformations, seizures, schizophrenia, and autism (Girirajan et al. 2011b; Girirajan and Eichler 2010). This variability in expressivity is thought to be due to additional rare events: the classical genetic modifier hypothesis (Girirajan et al. 2011b; Girirajan and Eichler 2010; Girirajan et al. 2011a, 2010; Pinto et al. 2010). Alternatively, such CNV lesions may not be causal but rather reflect decreased genome stability (Heng 2010a) and decreased robustness.

If decreased robustness correlates with or causes genome instability, patients with complex diseases should carry a higher burden of CNVs. This burden (or mutational load) may be inherited or may arise *de novo* through environmental stress in early development. Indeed, schizophrenia patients show a significantly higher global burden of rare CNVs (Walsh et al. 2008; International Schizophrenia Consortium 2008). Most importantly, private CNVs – that is, CNVs specific to a particular individual – are highly enriched in schizophrenia patients (International Schizophrenia Consortium 2008). A similar increase of CNV burden is found in autism patients (Pinto et al. 2010). Patients with the recurrent CNV on chromosome 16p12.1, which is associated with severe developmental delay, are also more likely to carry additional CNVs than matched controls (Girirajan et al. 2010). Moreover, patients with second CNV hits – possibly less robust individuals – show distinct and more severe clinical features (Girirajan et al.

2010). Consistent with decreased robustness, the facial symmetry of some patients is visibly perturbed (Girirajan et al. 2010). In nine other genomic disorders, additional CNV hits occur more frequently in patients than controls (Girirajan et al. 2010). This enrichment of CNV hits in patients is particularly strong for disorders with variable penetrance and expressivity (Girirajan et al. 2011b; Girirajan and Eichler 2010; Girirajan et al. 2010). In multiplex autism families (families with multiple occurrences), CNVs in affected siblings are fourfold enriched compared to unaffected siblings (Itsara et al. 2010). This observation led to the hypothesis that multiplex autism is due to an inherited predisposition in addition to other co-occurring mutations, including CNVs (Girirajan et al. 2011b; Itsara et al. 2010). We speculate that this predisposition may be decreased robustness.

The observation of additional, mostly private CNVs in patients is consistent with the existence of an extraordinarily large number of distinct genetic modifiers leading to disease. While certainly possible, this explanation is not consistent with our basic knowledge of genetic networks and their robustness to mutations. In yeast and worms, for which systematic analyses of single and double mutant phenotypes have been conducted, most loss-of-function mutations, even if combined, do not show a phenotype unless they occur in network hubs or master regulators such as HSP90 (Levy and Siegal 2008; Lehner et al. 2006; Boone et al. 2007; Davierwala et al. 2005; Giaever et al. 2002). The number of network hubs in humans is large but not infinitely large. Under our hypothesis, the increased CNV burden is an expression of a generally less robust and therefore sensitized background rather than a cause of disease. This explanation is consistent with the fact that the same CNV can be associated with different disorders. The different expressivity of a particular disorder may be due to different degrees of robustness loss or different revealed, formerly cryptic genetic variants in patients. In short, we propose that phenotypic robustness differs among humans as it does in model organisms and that those with lower robustness will be more susceptible to genetic and environmental perturbations and hence disease. This proposal is akin to assertions from physicians of generations past that certain people have a robust constitution whereas others have weak ones.

How can we assess robustness in humans to increase the heritability of complex diseases?

Currently, researchers attempt to identify unifying patterns of genetic variants, life style choices, and environmental factors in affected individuals (cases, **Fig. 4.2**). In contrast, we would

like to assess all individuals first for their level of phenotypic robustness (robust vs. not robust, **Fig. 4.2**). Individuals with significantly decreased robustness would then be analyzed for genetic variants associated with disease. This analysis will identify formerly cryptic variants that were unknown to influence disease risk. In addition, previously identified predisposing variants will increase in penetrance because healthy individuals carrying these variants will be robust and therefore not contribute to associations among the group with decreased robustness. In addition, this analysis will facilitate the identification of rare causal variants of large effect, which should be enriched in robust, but affected individuals. Increased penetrance of common variants together with revealed, formerly cryptic variants and causative rare alleles of large effect will significantly increase heritability for at least some complex diseases.

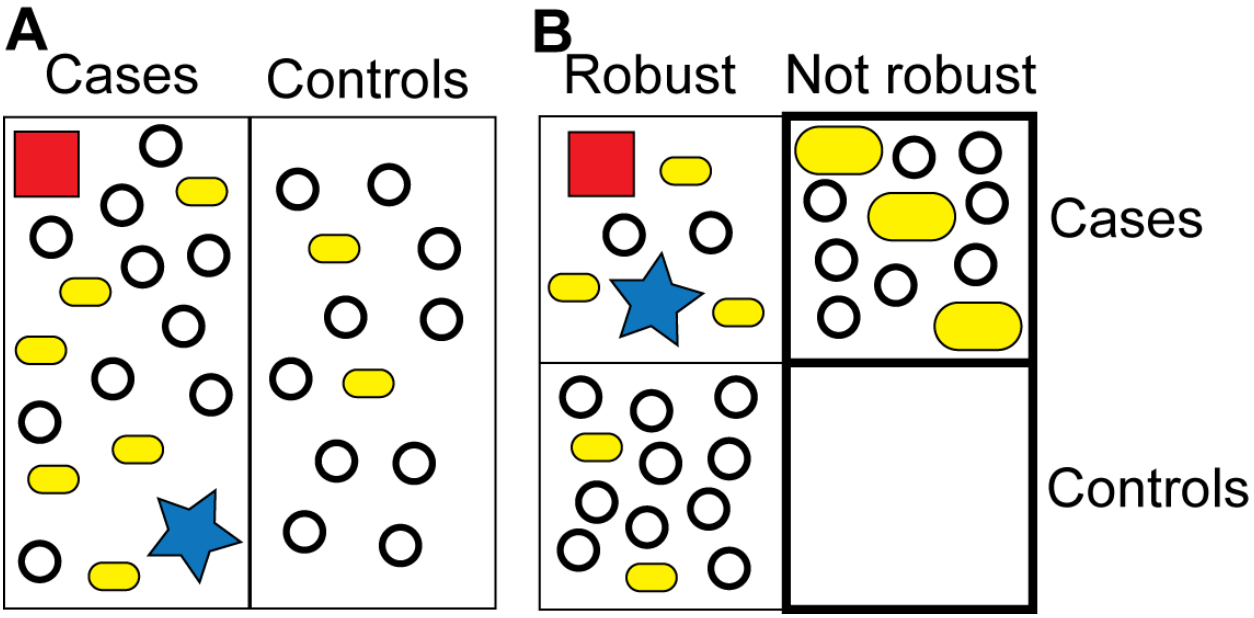

Figure 4.2. A. Current approach. GWAS identify variants that are overrepresented in cases. Rare variants of large effect (red square, blue star) may escape detection, thereby contributing to missing heritability. Common variants that are overrepresented in cases (small yellow bar, 6 versus 2) do not contribute strongly to disease risk. A cryptic disease-related variant does not show significant overrepresentation in cases (open circle). **B. Suggested approach.** Individuals are first analyzed for phenotypic robustness (bold box) and then for variants associated with disease. Rare variants of large effect will be enriched in robust cases, although they may also be present in non-robust cases. Variants that are overrepresented in all cases (robust, non-robust) will show higher penetrance in non-robust individuals (large yellow bars). The formerly cryptic, disease-related variant (open circle) is significantly enriched in non-robust cases versus s non-robust controls (and robust cases) and can therefore be identified. Together, heritability significantly increases.

The formerly cryptic genetic variant and higher penetrance variant can be thought of as "disease-specifiers" as they determine the specific disease phenotype of individuals carrying them. Note symbols represent highly simplified frequencies of specific variant in indicated groups and not individuals carrying certain variants.

This approach hinges critically on the identification of reliable markers for phenotypic robustness that can be readily assessed in large human populations. The functionality of master regulators such as HSP90 could potentially provide such a robustness measure. In yeast, however, the group of master regulators that affect robustness is large and functionally diverse (about 1-5% of all non-essential genes) (Levy and Siegal 2008). Assaying the diverse functions of hundreds of proteins is not a suitable high-throughput test.

Because decreased robustness correlates with and produces genome instability at several levels—an increase in microsatellite mutations, transposon mobility, recombination rates, base-substitution mutation rate, and large duplications and deletions—we suggest that these different genome instability events can serve as read-outs for decreased robustness. Recent technological advances have made most of these features easily accessible in large populations of humans and model organisms. The individual events may be rare, as observed for CNV variation; hence, they need to be investigated on a genome-wide scale. As microsatellites show by far the greatest mutation rate (Legendre et al. 2007), somatic microsatellite variation may be the most sensitive robustness marker. In fact, microsatellite variation has a long history as a marker for deficient DNA repair in certain cancers (Li 2008; Preston et al. 2010). At this point, however, assessment of microsatellite variation requires high-quality, expensive Sanger-sequencing. Neither the more cost-effective next-generation sequencing nor array-based genotyping can accurately determine variation of small microsatellites, but given the pace of current technology development, this technical hurdle should soon disappear.

Before investing in human studies, the suitability of molecular robustness markers is easily testable in model organisms. We envision the following proof-of-principle experiments: First, a diverse wild-type population, ideally freshly collected, is phenotyped for robustness, using traditional robustness measures such as symmetry or quantitative trait variation in genetically identical offspring (such as exists in plants and worms). We expect a distribution of robustness. Second, the least robust and most robust individuals are assessed for genome instability events at several levels. We expect to observe a higher frequency of these events in

less robust individuals. If a significant correlation between a traditional robustness measure and any genome instability event is found, this type of genome instability is a robustness marker that is applicable to humans. Third, if our hypothesis is correct, less robust individuals should be more susceptible to environmental stresses and will show higher expressivity of mutations and genetic variation. This assumption can be tested by exposing the least robust and most robust individuals to environmental stress and mutagenesis.

If genome instability events fail to predict robustness, alternatives exist. For humans, DNA-, RNA- or cell-based assays are preferable, as the necessary material can be obtained with relative ease. Indeed, several cell-based, high-throughput assays for somatic mutations already exist for humans (Albertini et al. 1993). In principle, these assays monitor allele loss resulting in an altered phenotypic output such as fluorescence. Another, potentially more promising, cell-based approach for determining robustness in humans would be to assess cell population variance for a given individual in gene expression, genome methylation, or chromatin states. Cell population variance in shape and other morphological features could also serve as a robustness marker (Levy and Siegal 2008). In yeast, this approach identified robustness master regulators by calculating the variance of 70 phenotypic traits among individual cells stained for nuclei, actin, and a cell wall marker (Levy and Siegal 2008). Given this arsenal of possible robustness measures, high-throughput molecular or cell-based robustness markers seem feasible in the near future.

Our hypothesis is testable even in the absence of human robustness markers. If decreased robustness predisposes to complex diseases, we can test whether individuals suffering from one complex disease are more likely to suffer from another, as we would predict. Another test would compare variants between cohorts of patients suffering from two different complex diseases. This comparison would filter out shared variants that are not causative and possibly related to decreased robustness and reveal statistically enriched causative variants that are specific to each disease.

Our hypothesis that robustness differences among individuals contribute to the missing heritability of disease is akin to prior propositions that epistasis, i.e. genetic interactions, accounts for the missing heritability (Gibson 2011; Zuk et al. 2012). Epistatically interacting loci could certainly be detected through GWAS, yet this will require very large sample sizes to find sufficient individuals of each genotype combination. In contrast, our approach reduces the

intractable complexity of possible genetic interactions to robustness, which we propose is a universal disease and trait modifier that can be feasibly measured in large human populations. Further, not all instances of disease may involve genetic interactions; some may arise from interactions of risk alleles with non-genetic mechanisms or environmental factors. Unlike traditional GWAS, our approach might capture these instances because robustness differences can arise through non-genetic mechanisms and environmental perturbations. Taking robustness into account has the potential to free us from disentangling the multitude of factors contributing to specific instances of disease. If successful, this approach might render complex disease more deterministic and predictable, allowing us to better identify the contributing life style choices and environmental exposures and ultimately decrease the severity and incidence of these devastating diseases.

Acknowledgements

This work was supported by grants from the National Human Genome Research Institute Interdisciplinary Training in Genomic Sciences (T32 HG00035) (KC) and the National Institute of Health (DP2OD008371) (CQ). We thank Elhanan Borenstein, Evan Eichler, Stanley Fields, Greg Gibson, Harmit Malik, Ray Monnat, and Maynard Olson for helpful discussions and comments. We thank Alexander Nuttle for sharing unpublished data.

CHAPTER FIVE

PREDICTING MUTATION PENETRANCE WITH MOLECULAR MARKERS OF ROBUSTNESS[4]

Introduction

Complex traits are phenotypes caused by interactions between multiple genetic loci, the environment, and stochastic noise. Complex traits, including complex human diseases such as Autism and Schizophrenia, cluster in families but cannot be completely explained by inheritance, leading researchers to conclude they have both genetic and environmental underpinnings. Genome-wide association studies have tried to identify the genetic underpinnings of complex disease and have been effective at identifying associated common variants, yet these loci only have modest explanatory value for the observed heritability. The inability to identify the genetic determinants underlying the heritable portion of complex traits has led to the term "missing heritability" (Eichler et al. 2010; Manolio et al. 2009). Many hypotheses have been put forth to account for the "missing heritability" of complex traits and some combination of these hypotheses together can likely explain most if not all of this phenomenon. One such hypothesis that others and we have put forth is that robustness, *i.e.* the ability to buffer genetic variants and environmental perturbations, varies between individuals and affects the penetrance of both standing variation and new mutations (Queitsch et al. 2012; Gibson 2009; Zuk et al. 2012). This would result in the same genetic variant having different effects in different individuals, preventing GWAS from discovering the causative genetic variant underlying a given complex trait.

The same phenomenon is also a potential mechanism allowing for rapid adaptation to new environment. For example, work on protein chaperone Heat Shock Protein 90 (HSP90) demonstrates that HSP90 can buffer genetic variation and that perturbation of HSP90 reveals previously cryptic genetic variation (Sangster et al. 2008b; Queitsch et al. 2002; Sangster et al. 2008a). This novel phenotypic variation can be acted upon by natural selection. There are other mechanisms thought to be able to buffer genetic variation such as redundancy of genetic networks and microRNAs (Wagner 2000; Bergman and Siegal 2003; Levy and Siegal 2008; Leclerc 2008; Ciliberti et al. 2007; Li et al. 2009; Posadas and Carthew 2014; Ito et al. 2011; Hornstein and Shomron 2006). When discussing robustness mechanisms in relationship to rapid adaptation, we invoke an environmental change or stress as the cause of the robustness

[4] This work is in collaboration with G. Alex Mason.

perturbation. For example, we know that heat stress can sequester HSP90 from its normal client base and reveal the same cryptic genetic variation revealed by specific reduction of HSP90 by drug treatment (Queitsch et al. 2002). When discussing the role robustness plays in "missing heritability," we instead invoke natural variation in the level of robustness between individuals. We have demonstrated natural variation of robustness across global and regional strains of Arabidopsis *thaliana* using the variance in quantitative traits across isogenic siblings as a measure of robustness for a given genotype. This measure of robustness is not an option in humans and is laborious in model organisms. To demonstrate robustness varies naturally in human populations, a new method to measure robustness is needed.

We propose genome stability as a molecular marker of robustness. There are multiple observations in the literature that in several organisms both HSP90 reduction and environmental stress lead to a decrease in genome stability (Mittelman et al. 2010; Specchia et al. 2010a; Chen et al. 2012; Queitsch et al. 2012). Here, we show that HSP90 reduction decreases genome stability in *Arabidopsis thaliana* using two approaches, a somatic homologous recombination reporter system (Schuermann et al. 2009) and our high-throughput, sequencing method MIPSTR. MIPSTR can detect somatic STR slippage, allowing for detection of subtle differences in genome stability. Unlike expression reporter systems, MIPSTR can measure genome stability in any organism with a high quality reference genome, including humans. We hypothesize that genome instability is an affect of a non-robust state, not a cause of phenotypic variation. To address this, we measured robustness using variation in quantitative traits in several well-characterized, genome stability mutants. We found that low genome stability does not necessarily translate to low robustness, validating our use of genome stability as a molecular marker for robustness.

With a molecular marker of robustness in hand, one can take robustness into account when associating genetic variants with complex traits. In our model of the role of robustness in complex traits, rare variants of large effect will underlie complex disease in robust individuals whereas both rare variants of large effect *and* unbuffered genetic variation, both standing variation and new mutations, will underlie complex traits in non-robust individuals. Work on HSP90 has demonstrated its ability to buffer standing genetic variation, variation that has been vetted over time by evolution. We hypothesize that HSP90 and other robustness regulators can

also buffer the effects of new mutations, such as recent, rare mutations or de novo mutations underlying some complex traits.

Here, in addition to characterizing nautral variation in robustness across diverse strains of *A. thaliana* and demonstrating a correlation between robustness and genome stability, we show that HSP90 can buffer new mutations. We show this with mutagenesis of *A. thaliana* seeds and the subsequent measurement of penetrance of mutations in HSP90 reduced and control plants. We follow up these results with a number of proposed future experiments to further test our hypothesis of the role of robustness in "missing heritability." First, we would like to associate genome stability with robustness in other robustness mutants and across diverse, non-mutant backgrounds. Second, we would like to verify our mutation penetrance measurements with a follow-up experiment. Third, we would like to correlate mutation penetrance with robustness and genome stability, again in non-robust mutants and natural strains. These experiments combined with the results presented here could be strong support for the role of robustness in both the "missing heritability" of complex traits and rapid adaptation.

Results

Measuring variance of quantitative traits confirmed natural variation of robustness

HSP90, our primary example of a robustness mechanism, not only buffers genetic variation but also buffers the effects of stochastic noise. Because of this, we use the variance in a quantitative trait across isogenic siblings as a measure of robustness (Sangster et al. 2008b). Those backgrounds that are robust will have less variation between siblings than those backgrounds that are not robust. Our lab previously identified variable levels of robustness, based on variance of hypocotyl length in isogenic siblings, across recombinant inbred lines (RILs) between two divergent *A. thaliana* strains Bay-0 (Bay) and Shadara (Sha). We reproduced the most and least robust Bay X Sha RILs across 70 individuals using the same robustness measure (**Fig 5.1** top), demonstrating reproducible natural variation of robustness in these strains.

We also found variation in robustness with this same measure in German strains that were collected recently from around Tubingen (Bomblies et al. 2010). These strains were genotyped with SNPs to determine diversity. We found that plants from a stand with low genetic diversity had very similar levels of robustness while plants from a stand with high levels of

genetic diversity had varied levels of robustness. This indicates that natural genetic variation plays a role in the natural variation of robustness (**Fig 5.1** middle, bottom).

Lastly, my labmate Jennifer Lachowiec, measured variance in hypocotyl and root length across 70 isogenic siblings for 96 diverse global strains of *A. thaliana* in triplicate. She found this robustness measure to be reproducible across replicates and to be highly variable between strains. She was in fact able to associate a SNP with variance in root length among these populations (Jennifer Lachowiec, personal communication), potentially identifying a new robustness master regulator with functional natural variants in a global population.

HSP90-reduced, non-robust plants have low genome stability

We first tested whether HSP90-reduced plants have higher rates of somatic homologous recombination (lower genome stability) than controls using a transgene-based homologous recombination reporter line. Error-free somatic homolgous recombination restores the function of the firefly luciferase (LUC) gene (Schuermann et al. 2009). We reduced HSP90 pharmacologically with geldenamycin (GDA), a drug that specifically inhibits HSP90. We grew reporter seedlings on media with and without GDA. We sprayed two-week-old seedlings with luciferin, the luciferase enzyme substrate, and counted luciferase-expressing spots, each of which represents a somatic homologous recombination event. There are more somatic events in HSP90-reduced seedlings than in controls (**Fig. 5.2**). Having confirmed that HSP90 reduction decreases genome stability as we previously observed with a similar GUS reporter system (**Fig. 4.1**), we then applied a more universally applicable method to address genome stability.

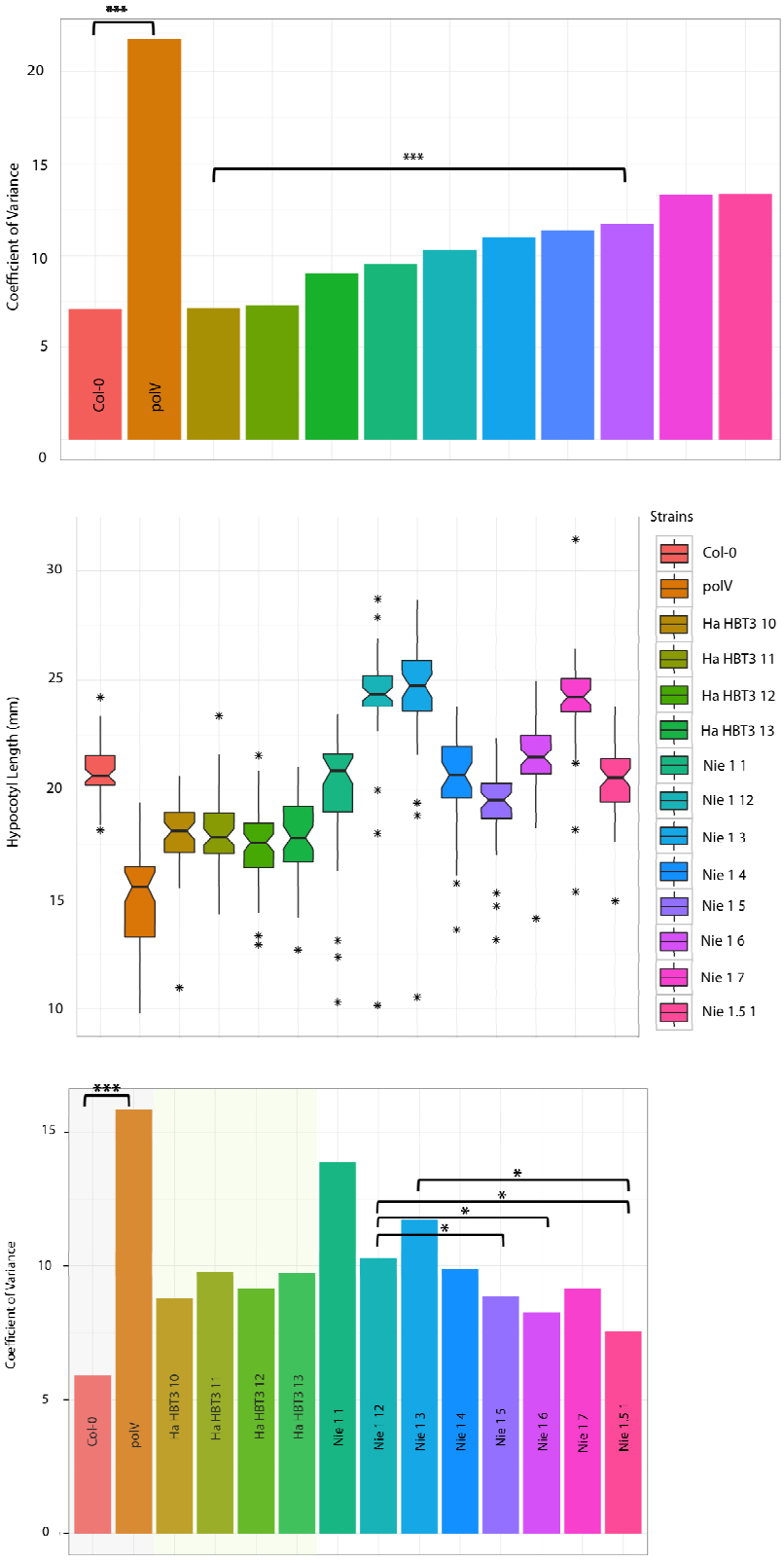

Figure 5.1 Variance in hypocotyl length as a measure of robustness in different genetic backgrounds. *p-value < 0.5, ***p-value < 0.005 in Levene's test comparison . Left-most genotype is Col-0 robust wild-type. **Top)** Coefficient of variance of hypocotyl length of different Bay X Sha RILs. There is significant natural variation among RILs, but none have a CV close to the robustness mutant positive control in orange (*polV*). **Middle)** Box-and-whiskers plot of hypocotyl length in regionally collected German strains. The Ha individuals have low genotypic diversity while the Nie individuals have high genetic diversity. **Bottom)** Coefficient of variance of hypocotyl length in regionally collected German strains. The diverse individuals vary in robustness while the non-diverse individuals do not. None of the natural strains have a CV close the robustness mutant positive control in orange (*polV*).

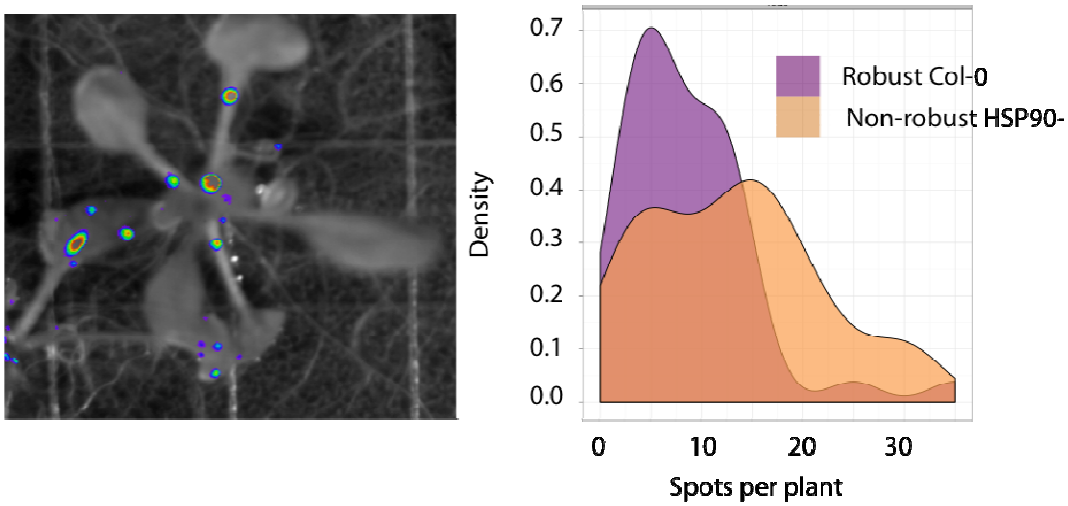

Figure 5.2 Non-robust plants have undergo more somatic homologous recombination events. Left) Photo of a 14-day-old seedling with glowing spots, each indicating a somatic homologous recombination event that completes a functional copy of the firefly luciferase gene. Right) Non-robust plants grown on GDA show more spots per plant than controls. Wilcoxon, p=0.011.

Applying our sequencing based method MIPSTR, we measured genome instability of short tandem repeats (*i.e.* microsatellite instability), in HSP90-RNAi lines and Col-0 wild-type control plants. With MIPSTR, we can identify the number of somatic mutations at ~90 tri- and hexa- nucleotide STR loci across the genome. In order to compare somatic alleles in HSP90-RNAi lines and wild-type controls, we performed a bootstrap resampling method to account for uneven sequencing library coverage. Previously, we used this method to assess genome instability in *Atmsh2* mutants, which are defective in their DNA mismatch repair pathway. *Atmsh2* mutants showed a ~5% increase in average somatic alleles per locus compared to Col-0 (**Supp. Fig. 3.4**). We expected a subtler increase in somatic alleles per locus in the HSP90-reduced plants. Comparing 39 loci across the genome that passed our read count threshold, we observed a 3.4% increase in average number of somatic alleles per locus in HSP90-RNAi-C1 plants than Col-0 controls (**Fig 5.3**, left).

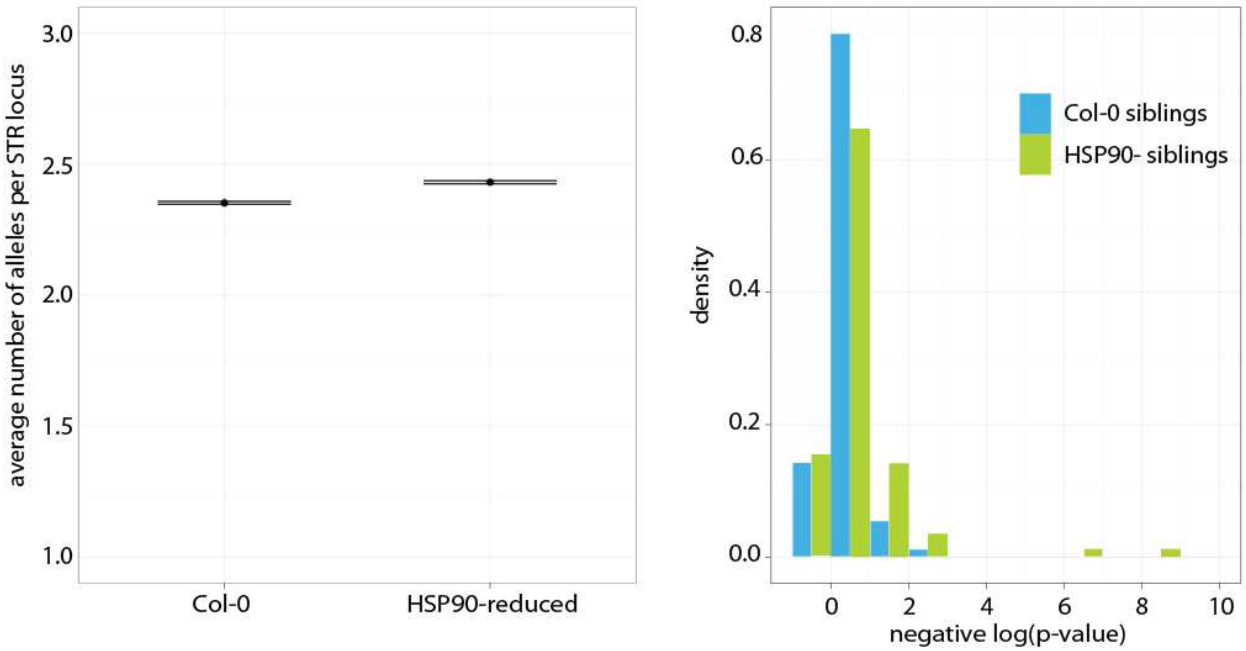

Figure 5.3 HSP90-RNAi plants show lower genome stability than Col-0 controls. Left) X-axis, Col-0 control or HSP90-RNAi-C1 plants. Y-axis, average number of alleles per STR locus (no somatic variation in homozygous plants equals one allele). Standard error is indicated with black lines. To calculate the average number of alleles per locus, we re-sampled 300 tag-defined read group modes at each STR locus 1000 times. The number of alleles supported by at least one tag-defined read group per sample was averaged across STR loci. **Right)** We used Fisher's exact test to ask whether the distribution of somatic alleles was different at each locus between sibling pairs. The negative $\log_{10}$ of the p-values are plotted for Col-0 siblings in blue and HSP90-RNAi-C1 siblings in green.

We also used MIPSTR to compare somatic differences between Col-0 siblings and HSP90-reduced siblings. If HSP90-reduced plants experience more somatic events than controls, we would expect the distribution of somatic alleles to show greater difference between HSP90 reduced siblings than between wild-type siblings. We compared the distribution of molecules (tag-defined read groups, Chapter 3) supporting different STR unit number genotypes across loci between HSP90 reduced siblings and Col-0 siblings using Fisher's exact test. HSP90-reduced siblings show more differences from each other somatically than Col-0 siblings, indicating a lower level of genome stability in the HSP90 reduced individuals than in controls (**Fig 5.3**, right).

Genome instability mutants do not necessarily have low robustness

To explore the relationship between genome stability and robustness, we measured robustness of the genome stability mutants *Atmsh2*, *Atpms1*, *Atwhy1*, *Atwhy3*, and two alleles of *Atpolδ1*. We expected that mutants in genes important for DNA stability would not necessarily have low robustness as in our hypothesis, genome instability is a consequence of low robustness and not a cause. *Atmsh2* mutants which have defects in mismatch repair have previously been shown to have high levels of somatic STR variation by others and us (Golubov et al. 2010a, **Supplemental Fig. 3.4)**. Mutants in *PMS1*, an important gene in the mismatch repair pathway and MutL homologue, show increased somatic STR variation and homologous recombination (Alou et al. 2004; Li 2008). Mutants in the WHY1 and WHY3, two genes important for plastid genome stability show increased rates of illegitimate, repeat-mediated somatic recombination (Maréchal et al. 2009). Mutants in *POLδ1*, a polymerase important for lagging strand replication and replication fidelity, show increased levels of homolgous combination with LUC reporter line, with the *polδ1-3* allele having a stronger affect than *polδ1-2* (Schuermann et al. 2009). We measured the variance of hypocotyl length across isogenic siblings these genome stability mutants in three replicates. and concluded genome stability does not determine robustness (**Fig 5.4**).

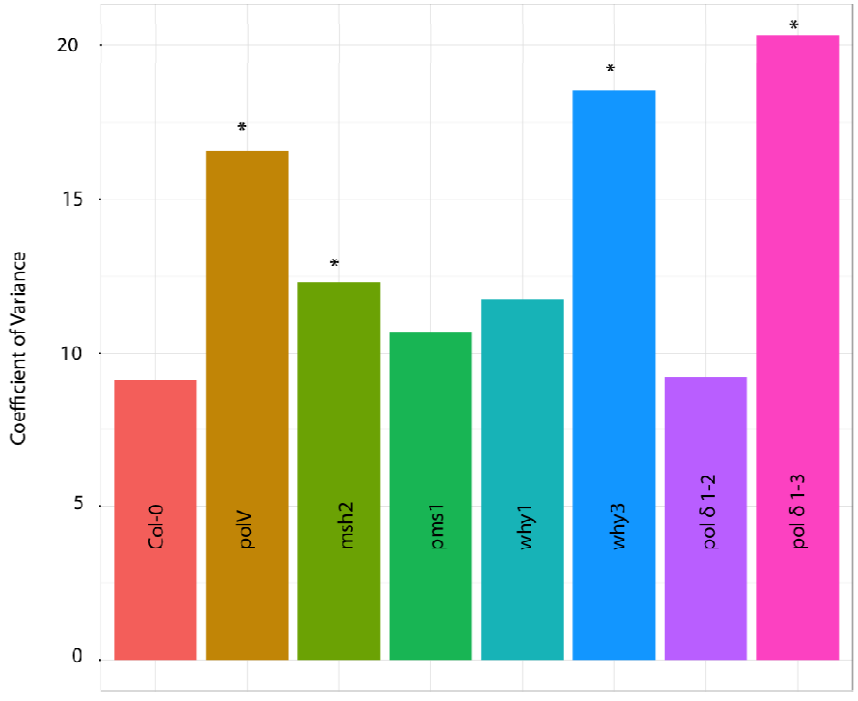

Figure 5.4 Variance in hypocotyl length as a measure of robustness in genome stability mutants. *p-value < 0.5 in a levene's test comparison with robust control Col-0 in salmon and non-robust mutant positive control *polV* in orange **Top)** Coefficient of variance of hypocotyl length of different genome stability mutants. All have decreased genome stability, but not all have increased robustness.

HSP90-reduced, non-robust plants have higher penetrance of new mutations

Next, we tested penetrance of new mutations in the Col-0 background. We mutagenized Col-0 seeds with two concentrations of EMS and mock treated control seeds. We planted seeds from each mutagenesis concentration and control seeds on both GDA-containing media and control media to measure penetrance of mutations in the M1 generation when the mutations are in the heterozygous state. We scored 10-day-old seedlings into five ranked scores ranging from wild-type (0) to very sick (4) (**Fig. 5.5A**). We measured penetrance by the statistical significance of the change from low scores to high scores after mutagenesis. The more significant a change meant higher mutation penetrance. In the M1 generation, we saw a large difference in penetrance of mutations in HSP90-reduced versus control seedlings (**Fig. 5.5B**). We performed the same experiment in the M2 generation, carrying forward only the lower concentration of EMS (see Methods) and again saw greater penetrance of new mutations in the HSP90-reduced plants than in controls, though this time the difference was much more subtle (**Fig. 5.5C**). We observed this increased penetrance whether we used bulked M2 seeds or M2 seeds pooled from a single silique from each plant. In the single silique seed pool, we increased representation of deleterious mutations by giving approximately equal representation to very affected and unaffected plants. In the bulk seed pool, healthy, unaffected plants produced many more seeds than affected plants and all seeds were collected. Because of this, we expected the single silique pool to be enriched for variants of large effect (fully penetrant variants regardless of the level of HSP90) and indeed, we see more extreme non-wild-type phenotypes in the single silique experiments compared to the bulked seed experiments in HSP90-reduced and control conditions (**Fig. 5.5D**). Throughout these experiments, we saw an affect of DMSO (the vehicle for GDA) on the germination of seedlings. To avoid using DMSO and to more directly compare our mutation penetrance results to our genome stability results, we mutagenized the HSP90-RNAi lines at the lower concentration of EMS. As with drug treatment, the HSP90-RNAi lines showed higher penetrance of mutations than controls in the M1 generation (**Fig. 5.5E**).

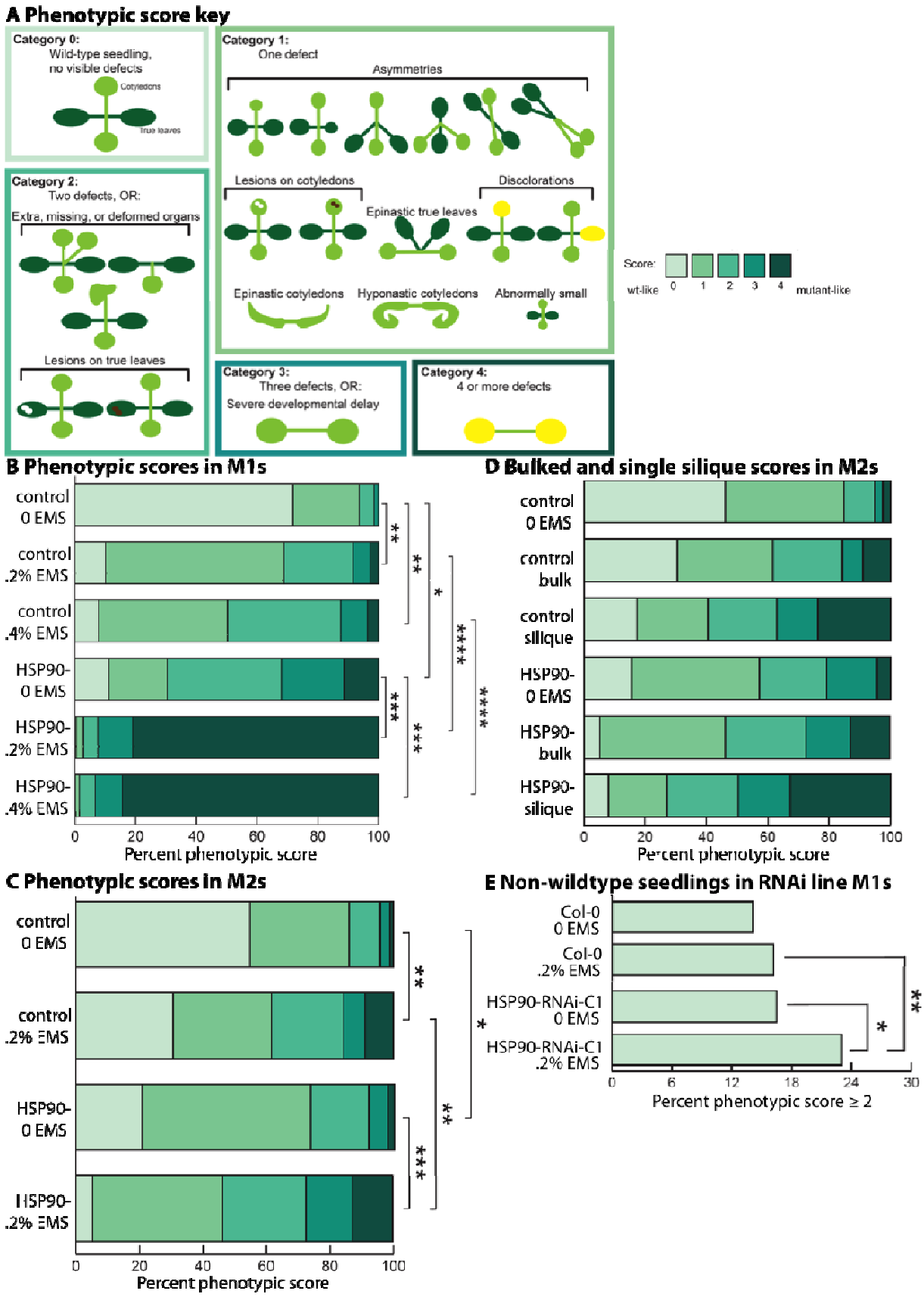
A Phenotypic score key
Category 0:
Wild-type seedling,
no visible defects
Cotyledons
True leaves
Category 1:
One defect
Asymmetries
Lesions on cotyledons
Discolorations
Epinastic true leaves
Epinastic cotyledons
Hyponastic cotyledons
Abnormally small
Category 2:
Two defects, OR:
Extra, missing, or deformed organs
Lesions on true leaves
Category 3:
Three defects, OR:
Severe developmental delay
Category 4:
4 or more defects
Score:
wt-like
0
1
2
3
4
mutant-like
B Phenotypic scores in M1s
control
0 EMS
control
.2% EMS
control
.4% EMS
HSP90-
0 EMS
HSP90-
.2% EMS
HSP90-
.4% EMS
0
20
40
60
80
100
Percent phenotypic score
D Bulked and single silique scores in M2s
control
0 EMS
control
bulk
control
silique
HSP90-
0 EMS
HSP90-
bulk
HSP90-
silique
0
20
40
60
80
100
Percent phenotypic score
C Phenotypic scores in M2s
control
0 EMS
control
.2% EMS
HSP90-
0 EMS
HSP90-
.2% EMS
0
20
40
60
80
100
Percent phenotypic score
E Non-wildtype seedlings in RNAi line M1s
Col-0
0 EMS
Col-0
.2% EMS
HSP90-RNAi-C1
0 EMS
HSP90-RNAi-C1
.2% EMS
0
6
12
18
24
30
Percent phenotypic score ≥ 2

Figure 5.5 HSP90-reduced plants show higher penetrance of mutation. A) Phenotypic score key from wild-type (0) to very sick (4). **B)** *p-value < 1E-6, **p-value < 1E-10, ***p-value <1E-50, ****p-value <1E-100 using proportional odds modeling. Phenotypic score distributions of M1 seedlings after mock EMS treatment or treatment with .2% or .4% EMS. Seedlings were grown on GDA (HSP90-) or control media for 10 days before scoring. **C)** *p-value < 1E-9, **p-value < 1E-10, ***p-value <1E-13 using proportional odds modeling. Phenotypic score distributions of M2 seedlings after mock EMS treatment or treatment with .2% EMS. Seedlings were grown on GDA (HSP90-) or control media for 10 days before scoring. **D)** Phenotypic score distirbutions of M2 seeds collected from a single silique of each plant (silique) versus bulked from all plants (bulk) after mock EMS treatment or treatment with .2% EMS. Seedlings were grown on GDA (HSP90-) or control media for 10 days before scoring. **E)** *p-value < .05, **p-value < 1E-10 from chi-square test. Percent of Col-0 Col-0 and HSP90-RNAi-C1 M1 seedlings after mock EMS treatment or treatment with .2% EMS with phenotypic score ≥ 2. Seedlings were then grown for 10 days before scoring.

Proposed Experiments

Does genome stability correlate with robustness in non-robust mutants other than HSP90-reduced and/or in natural populations?

We propose to apply a selected set of somatic variation sensitive MIPs and the MIPSTR method first to non-robust mutants and then to natural populations. Our lab has identified one other non-robust mutant in which we can test this hypothesis, although we do not yet know the mechanism by which the gene or gene product confers robustness (personal communication, Cris Alexandre). We will also test mutants in homologous genes to robustness regulators identified in other species, such as yeast and worms (Levy and Siegal 2008; Lehner et al. 2006). We will use this opportunity to test whether our robustness measures, namely variance of hypocotyl length among isogenic siblings and genome stability, correlate with other identifiers of robustness.

After confirming a correlation between STR stability and robustness in mutants, we will apply MIPSTR to natural populations for which we have measured robustness using variance of hypocotyl length, namely the Bay X Sha RILs, the regional German strains (Bomblies et al. 2010), and the global strains. We hypothesize that genome instability as measured by average number of alleles per locus correlates with variance in hypocotyl length among isogenic siblings in a given genetic background. If this is not the case, or the correlation is not perfect, we propose that genome stability itself could be an independent measure of robustness. Our lab has recently shown that some robustness master regulators likely only affect certain traits (Jen Lachowiec,

personal communication). Therefore, our hypocotyl variance measurement could miss certain types of robustness. Additionally, we have started to explore alternative robustness measures, such as global expression patterns from which we identify outliers as non-robust and "inliers" as robust. We will test whether this alternative measure correlates with hypocotyl length variance and/or genome stability in the Bay X Sha RILs on which the initial expression analysis has been completed (Kerry Bubb, personal communication).

Are we truly measuring mutation penetrance with our experimental design?

With our current experimental design, we do not directly show that the same mutation is more penetrant in HSP90-reduced plants than control plants. Instead, we show that the same degree of mutagenesis results in more non-wild-type phenotypes in HSP90 reduced plants than control plants. To address this, we will compare the same mutation in HSP90-reduced and control conditions. We selected 25 M2 seedlings from the following eight categories for transplanting to take to the next generation: 1) non-mutagenized, wild-type looking on control media 2) mutagenized, wild-type looking on control media 3) non-mutagenized, with a score of two, three, or four on control media 4) mutagenized, with a score of two, three, or four on control media 5) non-mutagenized, wild-type looking on HSP90 reduced media 6) mutagenized, wild-type looking on HSP90 reduced media 7) non-mutagenized, with a score of two, three, or four on HSP90-reduced media 8) mutagenized, with a score of two, three, or four on HSP90 reduced media (Fig. 5.6). We recorded and photographed the specific phenotypes we observed before transplanting.

Because *A. thaliana* is inbreeding, any mutations that were homozygous and affecting phenotype in the M2 generation will remain homozygous in the next generation. We will collect seeds from the transplants and will plant 72 seedlings of each genotype on control and HSP90-reduced media. Category one seedlings are unmutagenized and therefore should be isogenic, so we expect them to remain wild-type on control media and to show a slight increase in non-wild-type phenotypes on HSP90 reduced media due to reduced buffering of stochastic noise. Category two seedlings were mutagenized but we did not see an effect of mutations on phenotype, either because the mutations were neutral or because they were buffered. We expect category two seedlings to mostly appear wild-type on control media and to potentially show a specific non-wild-type phenotype on HSP90-reduced media if a previously buffered mutation is revealed. Category three seedlings are the result of stochastic noise, so we expect a low level of stochastic

noise on control media and a higher level of stochastic noise on HSP90-reduced media. Overall, we expect the results for category three to be similar to the results for category 1. Category four seedlings represent highly penetrant mutations of large effect. We expect to see the phenotype for which the parent seedling was originally selected among the seedlings growing on the control media and the same phenotype potentially at a higher frequency on HSP90-reduced media. Category five seedlings overcame the stochastic noise that resulted from the reduction of HSP90 during their development. We expect these seedlings to look mostly wild-type on control media with a subtly increased frequency of non-wild-type phenotypes on HSP90-reduced media. Category 6 seedlings either harbor neutral mutations or mutations that are lowly penetrant even upon HSP90 reduction. If the mutations turn out to be the latter, we would expect to see a novel phenotype on the HSP90-reduced media only and at a low frequency. We propose category 7 plants experienced high levels of stochastic noise after the reduction of HSP90, so we expect these plants to look mostly wild-type in the next generation, with a higher frequency of non-wild-type phenotypes on HSP90-reduced media than control. In this case, we do not necessarily expect to see the same phenotype we saw in the parent. Lastly, for category eight plants, we expect to see the phenotype of the parent at high frequency on HSP90 reduced media and potentially at a lower frequency on control media.

This experiment has the potential to demonstrate that the same mutation is more penetrant in HSP90-reduced, non-robust conditions than in control conditions and that mutations that look wild-type or have low penetrance in control conditions can be revealed or show high penetrance in non-robust conditions. These results would give us confidence in our approach to measuring mutation penetrance when we apply our measure in other genetic backgrounds. On the other hand, if we do not see consistent results in this experiment, we will need other measures of mutation penetrance.

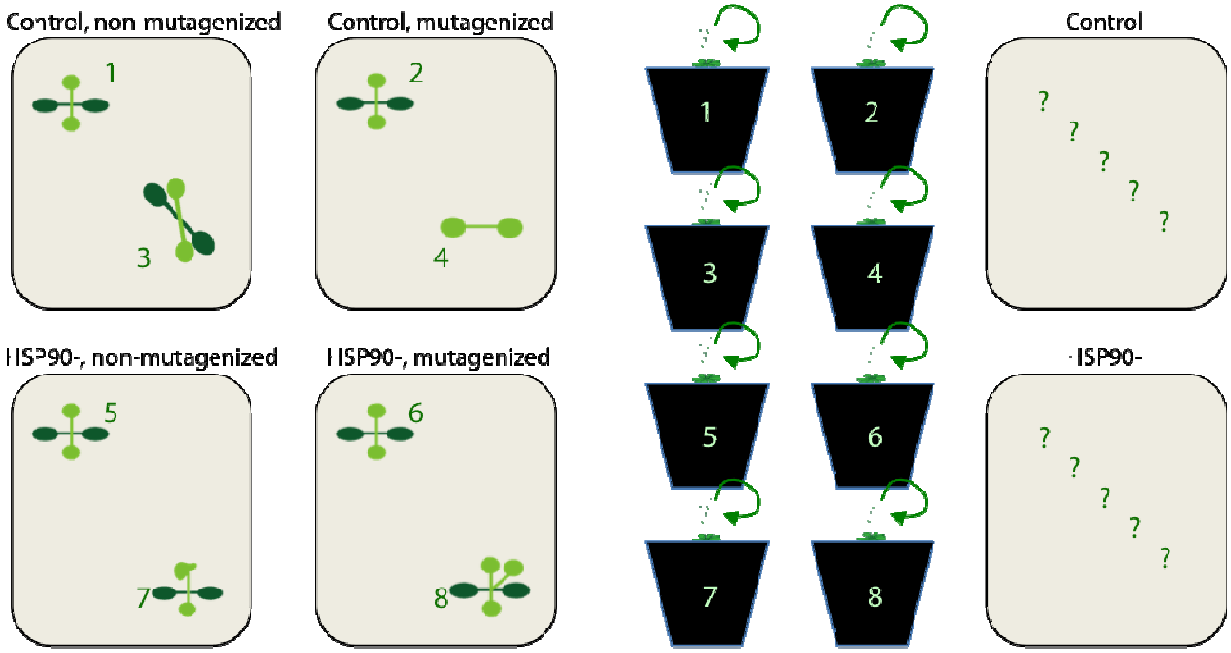

Figure 5.6 Proposed experiment to determine if a given mutation is more penetrant when HSP90 is reduced. From each of the eight categories shown on the four plates on the left, we chose 25 seedlings and recorded their phenotypes. We transplanted these seedlings to soil and will allow them to self-fertilize, leading to genetically identical offspring. We will plant seeds from each of these transplants on HSP90-reduced media and control media and record frequency of phenotypes.

Can we use quantitative traits as a measure of mutation penetrance?

If the previous set of experiments demonstrates that our mutation penetrance measure is not valid, we will use quantitative traits. As previously discussed, non-robust isogenic siblings have a higher variance in quantitative traits than robust, isogenic siblings. If mutation penetrance is higher in non-robust mutants, we expect the distribution of a quantitative trait in isogenic siblings to change more in non-robust plants than robust plants after mutagenesis. We can measure robustness as the significance in the change of the distributions, with the assumption that robust plants would experience a smaller shift in distribution than non-robust plants. This approach is more quantitative than our current approach and less prone to human bias. However, out of all the mutations, we can only expect some small percentage to affect a given complex trait. Because of this, it is possible that the seedlings without a mutation affecting the given trait with eclipse any small change in the distribution. For this reason, we believe our current

approach of taking into account all non-wild-type phenotypes is a more suitable measure of mutation penetrance.

Are we comparing apples to apples when comparing different mutagenized backgrounds?

In our experiments using GDA to reduce HSP90, we use the same batch of mutagenized Col-0 seeds, allowing us to control for the level of mutagenesis between the two conditions. In the case of the HSP90-RNAi-C1 line, other robustness mutants, and the natural populations, we can mutagenize them with the same protocol, but we cannot guarantee they experience the same number of mutations. This unequal mutagenesis can be due to variation in seed size or seed clumping among genetic backgrounds. In the case of the mutagenized RNAi-C1 line, we saw more non-wild-type phenotypes in the M1s (Fig 4sese) and more embryonic lethals in the M2s than in the mutagenized Col-0 controls, which could indicate either higher mutation penetrance as we proposed or the trivial explanation of more efficient mutagenesis in the HSP90-RNAi-C1 line. In order to make conclusions about mutation penetrance in these lines as compared to Col-0 or each other, we need to bulk and sequence the M2s to demonstrate that comparable amounts of mutations occurred in the different genetic backgrounds.

Do natural lines with low robustness and genome stability show high penetrance of mutations?

We will mutagenize natural lines from the Bay X Sha RIL population, the regional german population, and the global population from the extremes of robustness distributions derived from genome instability, variance in quantitative traits, expression outliers, or some combination of the three. We will measure both developmental phenotypes in M1s and developmental phenotypes and quantitative phenotypes in the M2 generation as described above. If we are able to predict mutation penetrance using a robustness molecular marker such as genome stability, we will be able to better understand which genetic variants are causative to complex traits.

Discussion

From the work presented here, we concluded that genome stability correlates with robustness in HSP90-reduced, non-robust plants and that penetrance of new mutations is increased in these plants. Although it has long been known that HSP90 reduction increases penetrance of or reveals standing genetic variation (Sangster et al. 2008b, 2008a; Queitsch et al.

2002), we showed for the first time that HSP90 reduction leads to increased penetrance of new mutations. We also demonstrated that robustness varies across natural populations using our traditional robustness measure of variance of a quantitative trait across isogenic siblings. With these results in hand, we can begin our proposed experiments to determine whether genome stability can serve as a measure of robustness in any genetic background and whether naturally low levels of robustness can lead to increased penetrance of new mutations. We have begun mutagenesis on highly and lowly robust Bay X Sha RILs.

The experimental approach presented here has a few caveats. The first is an incomplete understanding of the "right" measure of robustness. Our measure of variance of hypocotyl length across isogenic siblings could conceivably fail to identify robustness in other developmental stages or tissues. We know that variance of quantitative traits does not correlate well across traits (Jennifer Lachowiec, personal communication). We propose that genome stability will serve as a consistent marker of robustness and predictor of mutation penetrance, but this is yet to be shown.

Another caveat is the expectation that HSP90-reduction will always increase penetrance as opposed to generally changing penetrance, either increasing or decreasing it. When Sangster et al mapped HSP90 dependent loci in RILs, they observed new quantitative trait loci in HSP90-reduced conditions, but also saw QTLS disappear that were present in control RILs (Sangster et al. 2007, 2008a). This could be interpreted to mean that low robustness can either increase or decrease penetrance of genetic variation. The other, more plausible explanation for this is simple statistics. If revealed loci in the HSP90-reduced RILs had a much larger effect on phenotype than the loci identified in control RIls, the control QTLs would seem to "disappear." We will have a clear answer to this question from our proposed experiment to take specific mutations from wild-type and HSP90-reduced media to the next generation. If seedlings we selected as "wild-type" from HSP90 reduced conditions have a novel phenotype or a higher frequency of a phenotype on control media, we would conclude HSP90-reduction can also decrease penetrance of variants. This conclusion would void our mutation penetrance measurement as a valid approach to test for the effects of HSP90-reduction or low robustness.

There are alternative approaches to getting at the role of robustness in complex traits. Michael Dorrity in our lab is currently using deep mutational scanning (Araya and Fowler 2011) in genes of the mating pathway of *Saccharomyces cerevisae*, one of which is known to be an HSP90 client. We hypothesize that under control conditions, most mutations will be neutral but

under HSP90-reduced conditions some mutations will affect mating. We also predict large effect mutations that are fully penetrant and have an effect in both control and HSP90-redued conditions. Mike has already found variants that are buffered by HSP90 in a gene not known to code for an HSP90 client. His work can also address the question of whether HSP90-reduction always increases or can also decrease penetrance of new mutations. He is able to directly compare the same mutations in different conditions by using established techniques of pooling, barcoding and sequencing in yeast (Michael Dorrity, personal communication).

With the results presented here, the proposed experiments, and the complementary approaches currently addressing the same question, we are well on our way to understanding the role robustness plays in complex traits. With this knowledge in hand, we can begin to look in human populations for a way to stratify robust and non-robust humans, such as the application of MIPSTR to measure genome stability. Including robustness as an interaction in associations of genetic variants and phenotypes will likely reveal functional variants previously over looked because of their frequency in healthy, robust individuals.

Methods

Measuring robustness using hypocotyl length variance

Columbia-0 (Col-0) was used as wild type (WT). *Atdrd3-1* was used as a non-robust, positive control. All mutants were in the Columbia-0 background. Seedlings were grown vertically for seven days in the dark. Statistical significance of variance in hypocotyl length of 50-70 seedlings was determined using Levene's statistic.

Luciferase homologous recombination reporter line

Plants were grown for fourteen days on media with DMSO (mock) or geldanamycin dissolved in DMSO. At 14 days, seedlings were sprayed with 1uM Luciferin and kept in the dark for 1 hour before imaging on a NightOWL CCD camera. Spots were counted manually.

MIPSTR analysis

See Chapter 3 Methods.

MIPSTR to look at genome stability

To compare the number of somatic events occurring in different individuals, we only considered STR loci with low technical error scores (below 0.2, **Supplemental Table 1**) and with information for all plants in the comparison. We used bootstrap resampling to account for sometimes vastly different read counts. We resampled 300 modes of tag-defined read groups (See Chapter 3) for each locus. For each sample, we calculated the number of different STR unit number alleles present and averaged across loci.

EMS mutagenesis

3500 Col-0 seeds were pretreated with 0,005% Tween then mutagenized with .2% EMS, .4% EMS, and with a mock solution. Seeds rotated in EMS overnight. Seeds were thoroughly rinsed eight times, then suspended in agarose. 1000 seeds from each mutagenesis concentration were planted for M1 phenotyping. The remainder was planted to take to the M2 generation.

Developmental phenotyping

We scored 10-day-old seedlings according to the graphic in Figure 5. Only extreme asymmetries were considered to be different from wild-type.

CHAPTER 6

DISCUSSION AND FUTURE DIRECTIONS

The ability to rapidly adapt to new environments relies on phenotypic diversity on which natural selection can act. Genotypic variation as well as environment and stochastic noise underlie the phenotypic diversity. Often this genotype-phenotype map is not straightforward. For my dissertation, I investigated two molecular mechanisms of rapid adaptation. First, a class of highly mutable genetic elements called short tandem repeats. STRs are highly variable between individuals and reside in functional genomic regions, generating high levels of phenotypic diversity on which natural selection can act in a short time frame. Second, robustness master regulators can buffer genetic variation from phenotype, allowing it to accumulate without facing selective pressures, but when these robustness regulators are challenged by environmental stress or mutation, these previously cryptic genetic variants suddenly lead to novel and diverse phenotypes upon which natural selection can act.

The role of germline STR variation in adaptation and missing heritability

In the genotype-phenotype map, STRs are one class of genetic variation that has been largely overlooked simply due to the difficulty of accurately genotyping these loci with current whole genome re-sequencing data. Ignoring STR variation, however, will lead to holes in the genotype-phenotype map. This has been made evident by numerous, exhaustive studies on the function of specific STRs in a variety of organisms, which demonstrate a critical role of STR variation in phenotypic variation(Fondon and Garner 2004; Undurraga et al. 2012; Rival et al. 2014, 1; Rosas et al. 2014, 1; Smukalla et al. 2008, 1; Michael et al. 2007a; Peixoto et al. 1993). In order to fill in the map and to understand the role of STRs in rapid adaptation, the field needs accurate STR genotypes genome-wide across many individuals. To address this need, I developed a novel capture and sequencing based method called MIPSTR to accurately genotype STRs in a high-throughput manner.

In my proof-of-principle experiments, MIPSTR accurately called STR genotypes across the genomes of many individuals. With MIPSTR, I showed we can eliminate technical error from further analyses, accurately genotype heterozygous STR loci, and even confidently call somatic STR alleles. I showed that with only 100 loci, I could identify likely functional STRs and associate variation in STRs with phenotypic variation.

In the future, we will apply MIPSTR on a larger scale. In *A. thaliana*, we will target 2100 STR loci with unit lengths between 2-10 basepairs, including coding, regulatory, intronic, and matched intergenic loci. The matched intergenic loci are crucial as we hope to use them to build a neutral model of STR evolution. As is the case with most neutral models, this will not be perfect because even intergenic STR variation could have a phenotypic effect, for example on chromatin structure (Eckert and Hile 2009). This is also a particularly tricky question because STRs of different unit sequence, unit size, and total unit number mutate differently, so each potentially functional STR will have to be compared to the appropriate set of intergenic STRs in the neutral model (Gemayel et al. 2012).

With a neutral model in hand, we hope to build a phylogenetic tree of *A. thaliana* strains using genome-wide STR variation. STRs mutate faster than SNPs and therefore can be used to build a tree that has a finer time scale resolution. Having determined population structure and a neutral model, we could then identify STRs that are likely under selection and therefore functional. We will develop methods to robustly associate STR variation with phenotypic variation that take population structure into account. In particular, we will look for instances where STR variation can explain local adaptation to environment. We can then take candidate loci from these associations and functionally test the effects of STR unit number variation in relevant environmental conditions using transgenics (Undurraga et al. 2012).

MIPSTR can be applied to any organism with a high quality reference genome. My promising proof-of-principle results suggest that applying MIPSTR to humans, developing a robust neutral model, and applying sophisticated statistics for association will help to fill in the genotype-phenotype map, i.e. account for some “missing heritability” in complex disease. We know of several extreme cases where STR variation leads to disease phenotypes (Metzger et al. 2008; Gatchel and Zoghbi 2005; Zühlke et al. 2005), but we have little information on how subtler STR variation can affect human phenotype and disease. Because nearby SNPs and STRs are often not in LD (REF), STRs having a functional effect have likely been overlooked.

Because STRs are highly mutable, they serve as unique identifiers even in closely related individuals. The FBI uses just thirteen STR loci to uniquely identify individuals and match their DNA to that found at the scene of a crime (Rohlfs et al. 2012). Unfortunately, the set the FBI uses was derived from STRs that are highly variable in people of European ancestry, but these STRs are less variable and therefore less informative for other races (Rohlfs et al. 2012).

MIPSTR would allow the FBI to easily include more STR markers to remove the racial bias in their DNA matching protocol.

In order to demonstrate the role of STRs in rapid adaptation, I propose a laboratory evolution experiment using *Atmsh2* mutants, which are known to have a high rate of mutation in STRs (Golubov et al. 2010a). Mutation accumulation lines have been made in *A. thaliana*, but only SNPs and indels have been assessed (Ossowski et al. 2010). We could first use MIPSTR to assess STR variation among the mutation accumulation lines to understand STR mutation rate in a non-mutant background when not under selection. I propose to plant 36 *Atmsh2* mutant plants and 36 Col-0 controls, to grow these plants in an incubator on long days, and to rotate the plants daily. I would then apply a selection pressure for early flowering by selecting the six plants of each background that flower the earliest. I would collect and plant 36 seeds from each of these plants and again select the 6 earliest flowering from among the offspring of each genetic background. I would retain seeds from each generation for sequencing. After 10, 20, and 30 generations (2.5, 5, and 7.5 years respectively), I would apply MIPSTR and whole genome sequencing to identify variants under selection that may confer the early flowering phenotype. By doing this in the *Atmsh2* mutant background, I hope to enrich for functional STR variation. This experiment would definitively show the role of STR variation in adaptation.

The role of somatic STR variation in robustness, cancer, and development

I demonstrate that MIPSTR can accurately identify somatic STR variants in individuals. We propose to use this as a measure of genome stability and thereby robustness of organisms, but there are many other potential uses, the most obvious of which is to measure microsatellite instability (MSI) in tumors. Traditionally, researchers separately PCR amplify several STR loci in a tumor sample and matched normal sample and then use capillary electrophoresis to compare the sizes and distributions of STR alleles (Boland et al. 1998a; Kim et al. 2013). If the distribution is wider or shifted in the tumor, this tumor is said to have MSI (Boland et al. 1998a; Kim et al. 2013). This conclusion can affect diagnosis, prognosis, and treatment of the cancer (Boland et al. 1998a; Kim et al. 2013). MIPSTR can easily provide this same information and can do so at a much higher resolution allowing finer distinctions of the level of MSI. In fact, MIPSTR could potentially detect somatic variants very early on in tumor development. Similarly, MIPSTR could measure the genetic heterogeneity of STR alleles in a tumor. It is

thought that more heterogeneous tumors have higher evolutionary potential and are therefore more aggressive and more likely to develop resistance to therapeutics (Fox et al. 2013; Schmitt et al. 2012). MIPSTR could discern relative presence of somatic variants that are not present in the dominant clone, avoiding the need to perform single-cell sequencing analyses to determine tumor heterogeneity (Baslan et al. 2012; Navin et al. 2011).

There is evidence that somatic STR variation is not only seen in tumors as a result of a defective mismatch repair pathway, but that it also may play a crucial role in normal development. In particular, somatic STR variation has been proposed to play a role in neural development and neurodegenerative diseases (Poduri et al. 2013; Fehér 2014; Fondon et al. 2008). If this is the case, tracking somatic variants throughout development using an approach like MIPSTR could illucidate much more of the genotype-phenotype map than originally predicted.

A. thaliana would be an excellent model to address the role of somatic STR variation in development. To look at this, I would plant both Col-0 and *Atmsh2* plants. I would collect DNA from the plants at various life stages and tissues, ideally able to compare tissues and stages of the same plant if enough DNA can be extracted. I would apply MIPSTR to these DNA samples to ask if I see any consistent differences in specifc STR loci between tissues in Col-0 plants. I would look to see if these differences were disrupted in *Atmsh2* plants and if that disruption would affect phenotype in the respective tissue. The results of these experiments would be difficult to functionally test. I would like to design an inducible STR mutant, perhaps using heat as an inducer, but currently this type of transgenic *A. thaliana* is far off. If this experiment has reproducible results, this would be transformative in understanding how genotype affects phenotype and heritability.

The role of robustness in "missing heritability" and open questions

I put forth the hypothesis that robustness, or the ability of an organism to buffer genetic variation, is important for both rapid adaptation and understanding the "missing heritability" of complex traits. In order to address this hypothesis, we first need a method to measure the robustness of an individual. I suggested genome stability as a molecular marker and presented MIPSTR as a method to assess genome stability of individuals. So far, my work supports the

idea that genome stability can serve as a marker for robustness and that non-robust individuals show higher penetrance of new mutations. We plan to extend this analysis to natural strains.

Our data suggests we can apply MIPSTR to humans and measure genome instability of individuals. With this information in hand, we could begin to look at the interaction between robustness and genetic variants leading to complex disease. Current data implicates chromatin remodeling factors in both Schizophrenia and Autism, two very different neurological disorders (McCarthy et al. 2014; Walsh et al. 2008; O'Roak et al. 2012). This begs the question, why would disruption of such an important and general function cause two specific diseases? Instead, we could interpret these variants in chromatin remodeling factors as variants that lower robustness in individuals, potentially by disrupting chromatin structure and gene expression. If we consider only individuals with these disruptive mutations and separate by disease phenotype, we may be able to identify the variants that are neutral in robust individuals but are the underlying cause of the specific phenotypes in these patients.

After the preliminary experiments presented in Chapter 5, we are left with open questions like "how generalizable is the HSP90 example of a robustness mechanisms?" and "are there different types of robustness?" HSP90 affects many traits and many pathways, as is expected due to its molecular role as a protein chaperone helping many different classes of proteins fold and function (REF). From recent work in our lab, we expect there will also be pathway specific robustness mechanisms that like HSP90 can be perturbed so they reveal genetic variation or the effects of stochastic noise, but only in a given pathway or developmental stage. It has been suggested that some robustness mechanisms buffer either genetic variation or stochastic noise, but not both. Furthering understanding of types of robustness and what a given mechanism buffers will further our understanding of the role of robustness in heritability and adaptation.

To further this understanding, I could further address the question of causality concerning robustness and somatic variation. I would start by testing somatic stability of isogenic siblings using MIPSTR. In particular, I would ask if isogenic outliers in quantitative traits (e.g. those with either very short of very long hypocotyls) have higher levels of somatic variation. If this were the case, it would suggest that robustness to stochastic noise- or the ability to attain the trait mean- is also correlated to genome stability along with robustness to genetic variation. We can also use MIPSTR to better understand the heritability of robustness and genome stability. Because I see

differences in robustness in different genetic backgrounds and predict genome stability can serve as a molecular marker for this robustness, we would expect genome stability would be heritable across generations. Applying MIPSTR across generations of *A. thaliana* individuals would help us discern the comparative roles of heritability and stochastic noise in determining the robustness of an individual.

In my dissertation, I presented a novel method called MIPSTR to genotype STRs accurately, genome-wide, and across many individuals. This method can genotype germline and somatic STR variation taking into account the high levels of technical error. I began to apply this technology to understand the role of germline STR variation in rapid adaptation and "missing heritability." I also applied MIPSTR to measure genome stability of robust wild-type and non-robust HSP90-reduced plants and showed that new mutations are more penetrant in plants with low robustness and low genome stability. I also proposed a number of experiments applying MIPSTR to allow further understanding of the role of STRs and robustness in both adaptation and "missing heritability."

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Albertini RJ, Nicklas JA, Fuscoe JC, Skopek TR, Branda RF, O'Neill JP. 1993. In vivo mutations in human blood cells: biomarkers for molecular epidemiology. *Environ Health Perspect* **99**: 135–141.

Alou AH, Azaiez A, Jean M, Belzile FJ. 2004. Involvement of the Arabidopsis thaliana AtPMS1 gene in somatic repeat instability. *Plant Mol Biol* **56**: 339–349.

Araya CL, Fowler DM. 2011. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol* **29**: 435–442.

Armour JAL, Patel I, Thein SL, Fey MF, Jeffreys AJ. 1989. Analysis of somatic mutations at human minisatellite loci in tumors and cell lines. *Genomics* **4**: 328–334.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**: 627–631.

Baggs JE, Price TS, DiTacchio L, Panda S, Fitzgerald GA, Hogenesch JB. 2009. Network features of the mammalian circadian clock. *PLoS Biol* **7**: e52.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, et al. 2012. Genome-wide copy number analysis of single cells. *Nat Protoc* **7**: 1024–1041.

Bergman A, Siegal ML. 2003. Evolutionary capacitance as a general feature of complex gene networks. *Nature* **424**: 549–552.

Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. 2013. Finding the sources of missing heritability in a yeast cross. *Nature* **494**: 234–7.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, et al. 1998a. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* **58**: 5248–5257.

Bolton KA, Ross JP, Grice DM, Bowden NA, Holliday EG, Avery-Kiejda KA, Scott RJ. 2013. STaRRRT: a table of short tandem repeats in regulatory regions of the human genome. *BMC Genomics* **14**: 795.

Bomblies K, Yant L, Laitinen RA, Kim S-T, Hollister JD, Warthmann N, Fitz J, Weigel D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genet* **6**: e1000890.

Boone C, Bussey H, Andrews BJ. 2007. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**: 437–449.

Boyko A, Kovalchuk I. 2011. Genome instability and epigenetic modification--heritable responses to environmental stress? *Curr Opin Plant Biol* **14**: 260–266.

Burga A, Casanueva MO, Lehner B. 2011. Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* **480**: 250–253.

Butler AP, Trono D, Coletta L Della, Beard R, Fraijo R, Kazianis S, Nairn RS. 2007a. Regulation of CDKN2A/B and Retinoblastoma genes in Xiphophorus melanoma. *Comp*

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011a. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* **43**: 956–963.

Cao MD, Balasubramanian S, Bodén M. 2014a. Sequencing technologies and tools for short tandem repeat variation detection. *Brief Bioinform* bbu001–.

Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, Balasubramanian S, Bodén M. 2014b. Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res* **42**: e16.

Carey CC, Gorman KF, Rutherford S. 2006. Modularity and intrinsic evolvability of Hsp90-buffered change. *PloS One* **1**: e76.

Casanueva MO, Burga A, Lehner B. 2012. Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans. *Science* **335**: 82–85.

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A, et al. 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301**: 386–9.

Chapal-Ilani N, Maruvka YE, Spiro A, Reizel Y, Adar R, Shlush LI, Shapiro E. 2013. Comparing algorithms that reconstruct cell lineage trees utilizing information on microsatellite mutations. ed. W.W. Wasserman. *PLoS Comput Biol* **9**: e1003297.

Chen G, Bradford WD, Seidel CW, Li R. 2012. Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature* **482**: 246–250.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009a. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677–681.

Ciliberti S, Martin OC, Wagner A. 2007. Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A* **104**: 13591–13596.

Cook LM, Saccheri IJ. 2013. The peppered moth and industrial melanism: evolution of a natural selection case study. *Heredity* **110**: 207–212.

Cooley MB, Carychao D, Nguyen K, Whitehand L, Mandrell R. 2010. Effects of environmental stress on stability of tandem repeats in Escherichia coli O157:H7. *Appl Environ Microbiol* **76**: 3398–3400.

Davierwala AP, Haynes J, Li Z, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding H, Zhu H, Chen Y, et al. 2005. The synthetic genetic interaction spectrum of essential genes. *Nat Genet* **37**: 1147–1152.

De Visser JAGM, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, et al. 2003. Perspective: Evolution and detection of genetic robustness. *Evol Int J Org Evol* **57**: 1959–1972.

Debat V, David P. 2001. Mapping phenotypes: canalization, plasticity and developmental stability. *Trends Ecol Evol* **16**: 555–561.

Dittmar EL, Oakley CG, Agren J, Schemske DW. 2014. Flowering time QTL in natural populations of Arabidopsis thaliana and implications for their adaptive value. *Mol Ecol*.

Dote H, Burgan WE, Camphausen K, Tofilon PJ. 2006. Inhibition of hsp90 compromises the DNA damage response to radiation. *Cancer Res* **66**: 9211–9220.

Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, Verstrepen KJ, Froyen G. 2014a. Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res* gku212–.

Dworkin I. 2005. A study of canalization and developmental stability in the sternopleural bristle system of Drosophila melanogaster. *Evol Int J Org Evol* **59**: 1500–1509.

Eckert KA, Hile SE. 2009a. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog* **48**: 379–388.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**: 446–50.

Eisenegger C, Pedroni A, Rieskamp J, Zehnder C, Ebstein R, Fehr E, Knoch D. 2013. DAT1 polymorphism determines L-DOPA effects on learning about others' prosociality. ed. S.L. Sensi. *PloS One* **8**: e67820.

Escher D, Bodmer-Glavas M, Barberis A, Schaffner W. 2000. Conservation of Glutamine-Rich Transactivation Function between Yeast and Humans. *Mol Cell Biol* **20**: 2774–2782.

Fehér A. 2014. Somatic embryogenesis - Stress-induced remodeling of plant cell fate. *Biochim Biophys Acta*.

Fondon JW 3rd, Garner HR. 2004a. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101**: 18058–18063.

Fondon JW, Hammock EAD, Hannan AJ, King DG. 2008a. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci* **31**: 328–334.

Fox EJ, Prindle MJ, Loeb LA. 2013. Do mutator mutations fuel tumorigenesis? *Cancer Metastasis Rev* **32**: 353–361.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011a. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**: 419–423.

Gangaraju VK, Yin H, Weiner MM, Wang J, Huang XA, Lin H. 2011. Drosophila Piwi functions in Hsp90-mediated suppression of phenotypic variation. *Nat Genet* **43**: 153–158.

Gangestad SW, Thornhill R, Garver-Apgar CE. 2005. Women's sexual interests across the ovulatory cycle depend on primary partner developmental instability. *Proc Biol Sci* **272**: 2023–2027.

Gatchel JR, Zoghbi HY. 2005a. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet* **6**: 743–755.

Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. 2012a. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes* **3**: 461–480.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**: 445–77.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**: 387–391.

Gibson G. 2009. Decanalization and the origin of complex disease. *Nat Rev Genet* **10**: 134–140.

Gibson G. 2011. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**: 135–45.

Gidalevitz T, Wang N, Deravaj T, Alexander-Floyd J, Morimoto RI. 2013. Natural genetic variation determines susceptibility to aggregation or toxicity in a C. elegans model for polyglutamine disease. *BMC Biol* **11**: 100.

Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M, et al. 2011a. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet* **7**: e1002334.

Girirajan S, Campbell CD, Eichler EE. 2011b. Human copy number variation and complex genetic disease. *Annu Rev Genet* **45**: 203–226.

Girirajan S, Eichler EE. 2010. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19**: R176–187.

Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L, Walsh T, McCarthy SE, Baker C, et al. 2010. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42**: 203–209.

Golubov A, Yao Y, Maheshwari P, Bilichak A, Boyko A, Belzile F, Kovalchuk I. 2010a. Microsatellite instability in Arabidopsis increases with plant development. *Plant Physiol* **154**: 1415–1427.

Gordon M. 1927. THE GENETICS OF A VIVIPAROUS TOP-MINNOW PLATYPOECILUS; THE INHERITANCE OF TWO KINDS OF MELANOPHORES. *Genetics* **12**: 253–283.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–22.

Greenblum S, Turnbaugh PJ, Borenstein E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* **109**: 594–599.

Grimm D, Hagmann J, Koenig D, Weigel D, Borgwardt K. 2013a. Accurate indel prediction using paired-end short reads. *BMC Genomics* **14**: 132.

Gruber JD, Vogel K, Kalay G, Wittkopp PJ. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects, and dominance. *PLoS Genet* **8**: e1002497.

Guilmatre A, Highnam G, Borel C, Mittelman D, Sharp AJ. 2013a. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum Mutat* **34**: 1304–1311.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012a. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**: 1154–1162.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2010a. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinforma Oxf Engl* **26**: 1277–1283.

Hammock EAD, Young LJ. 2005a. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**: 1630–1634.

Hannan AJ. 2010. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for “missing heritability”. *Trends Genet* **26**: 59–65.

Hayes B. 2013a. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Methods Mol Biol Clifton NJ* **1019**: 149–169.

Heng HHQ. 2010a. Missing heritability and stochastic genome alterations. *Nat Rev Genet* **11**: 813.

Hiatt JB, Pritchard CC, Salipante SJ, O’Roak BJ, Shendure J. 2013a. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23**: 843–854.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013a. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**: e32.

Hornstein E, Shomron N. 2006. Canalization of development by microRNAs. *Nat Genet* **38 Suppl**: S20–24.

Hugot J-P, Alberti C, Berrebi D, Bingen E, Cézard J-P. 2003. Crohn’s disease: the cold chain hypothesis. *Lancet* **362**: 2012–2015.

Inanir A, Tural S, Yigit S, Kalkan G, Pancar GS, Demir HD, Ates O. 2013. Association of IL-4 gene VNTR variant with deep venous thrombosis in Behçet’s disease and its effect on ocular involvement. *Mol Vis* **19**: 675–683.

International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**: 237–241.

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**: 115–119.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. De novo rates and selection of large copy number variation. *Genome Res* **20**: 1469–1481.

Jarosz DF, Lindquist S. 2010. Hsp90 and environmental stress transform the adaptive value of natural genetic variation. *Science* **330**: 1820–1824.

Jarosz DF, Taipale M, Lindquist S. 2010. Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. *Annu Rev Genet* **44**: 189–216.

Kashi Y, King D, Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* **13**: 74–78.

Kim T-M, Laird PW, Park PJ. 2013a. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155**: 858–868.

King DG. 2012. Indirect selection of implicit mutation protocols. *Ann N Y Acad Sci* **1267**: 45–52.

Kosswig C. 1928. Über Kreuzungen zwischen den Teleostiern Xiphophorus helleri und Platypoecilus maculatus. *Z Indukt Abstamm-Vererbungsl* **47**: 150–158.

Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M, Hohn B. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature* **423**: 760–2.

Lachowiec J, Lemus T, Borenstein E, Queitsch C. 2014. *Hsp90 promotes kinase evolution*. http://biorxiv.org/lookup/doi/10.1101/006411 (Accessed August 20, 2014).

Lachowiec J, Lemus T, Thomas JH, Murphy PJM, Nemhauser JL, Queitsch C. 2013. The protein chaperone HSP90 can facilitate the divergence of gene duplicates. *Genetics* **193**: 1269–1277.

Laidlaw J, Gelfand Y, Ng K-W, Garner HR, Ranganathan R, Benson G, Fondon JW. 2007. Elevated Basal Slippage Mutation Rates among the Canidae. *J Hered* **98**: 452–460.

Leclerc RD. 2008. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol* **4**: 213.

Lee HS, Lee BL, Kim SH, Woo DK, Kim HS, Kim WH. 2001. Microsatellite instability in synchronous gastric carcinomas. *Int J Cancer J Int Cancer* **91**: 619–24.

Lee J-M, Zhang J, Su AI, Walker JR, Wiltshire T, Kang K, Dragileva E, Gillis T, Lopez ET, Boily M-J, et al. 2010. A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst Biol* **4**: 29.

Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007a. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res* **17**: 1787–1796.

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* **4**: 170.

Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. 2006. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* **38**: 896–903.

Levy SF, Siegal ML. 2008. Network hubs buffer environmental variation in Saccharomyces cerevisiae. *PLoS Biol* **6**: e264.

Lewandowska-Sabat AM, Winge P, Fjellheim S, Dørum G, Bones AM, Rognli OA. 2012. Genome wide transcriptional profiling of acclimation to photoperiod in high-latitude accessions of Arabidopsis thaliana. *Plant Sci Int J Exp Plant Biol* **185-186**: 143–155.

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al. 2008a. Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.

Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008b. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**: 776–788.

Li G-M. 2008. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**: 85–98.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* **25**: 1754–1760.

Li H, Liu J, Wu K, Chen Y. 2012. Insight into role of selection in the evolution of polyglutamine tracts in humans. ed. R.A. Veitia. *PloS One* **7**: e41167.

Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. 2009. A MicroRNA Imparts Robustness against Environmental Fluctuation during Development. *Cell* **137**: 273–282.

Mackay TFC. 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet* **15**: 22–33.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012a. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Malekzadeh F, Alberti C, Nouraei M, Vahedi H, Zaccaria I, Meinzer U, Nasseri-Moghaddam S, Sotoudehmanesh R, Momenzadeh S, Khaleghnejad R, et al. 2009. Crohn's disease and early exposure to domestic refrigeration. *PloS One* **4**: e4288.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–53.

Manu null, Surkova S, Spirov AV, Gursky VV, Janssens H, Kim A-R, Radulescu O, Vanario-Alonso CE, Sharp DH, Samsonova M, et al. 2009. Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors. *PLoS Comput Biol* **5**: e1000303.

Maréchal A, Parent J-S, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. 2009. Whirly proteins maintain plastid genome stability in Arabidopsis. *Proc Natl Acad Sci U S A* **106**: 14693–14698.

Masel J, Siegal ML. 2009. Robustness: mechanisms and consequences. *Trends Genet TIG* **25**: 395–403.

McCarthy SE, Gillis J, Kramer M, Lihm J, Yoon S, Berstein Y, Mistry M, Pavlidis P, Solomon R, Ghiban E, et al. 2014. De novo mutations in schizophrenia implicate chromatin

remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry* **19**: 652–658.

Meierjohann S, Schartl M. 2006. From Mendelian to molecular genetics: the Xiphophorus melanoma model. *Trends Genet TIG* **22**: 654–61.

Metzger S, Rong J, Nguyen H-P, Cape A, Tomiuk J, Soehn AS, Propping P, Freudenberg-Hua Y, Freudenberg J, Tong L, et al. 2008. Huntingtin-associated protein-1 is a modifier of the age-at-onset of Huntington's disease. *Hum Mol Genet* **17**: 1137–46.

Michael TP, Park S, Kim T-S, Booth J, Byer A, Sun Q, Chory J, Lee K. 2007a. Simple sequence repeats provide a substrate for phenotypic variation in the Neurospora crassa circadian clock. *PloS One* **2**: e795.

Mittelman D, Sykoudis K, Hersh M, Lin Y, Wilson JH. 2010. Hsp90 modulates CAG repeat instability in human cells. *Cell Stress Chaperones* **15**: 753–9.

Molinier J, Ries G, Zipfel C, Hohn B. 2006. Transgeneration memory of stress in plants. *Nature* **442**: 1046–1049.

Molla M, Delcher A, Sunyaev S, Cantor C, Kasif S. 2009. Triplet repeat length bias and variation in the human transcriptome. *Proc Natl Acad Sci U S A* **106**: 17095–17100.

Morrison NA, Stephens AA, Osato M, Polly P, Tan TC, Yamashita N, Doecke JD, Pasco J, Fozzard N, Jones G, et al. 2012. Glutamine repeat variants in human RUNX2 associated with decreased femoral neck BMD, broadband ultrasound attenuation and target gene transactivation. ed. E.G. Laird. *PloS One* **7**: e42617.

Morrison NA, Stephens AS, Osato M, Pasco JA, Fozzard N, Stein GS, Polly P, Griffiths LR, Nicholson GC. 2013. Polyalanine repeat polymorphism in RUNX2 is associated with site-specific fracture in post-menopausal females. ed. S.F.A. Grant. *PloS One* **8**: e72740.

Mularoni L, Guigó R, Albà MM. 2006a. Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biol* **7**: R33.

Mularoni L, Ledda A, Toll-Riera M, Albà MM. 2010. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res* **20**: 745–54.

Mularoni L, Veitia RA, Albà MM. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89**: 316–25.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472**: 90–94.

Nelson RM, Pettersson ME, Carlborg Ö. 2013. A century after Fisher: time for a new paradigm in quantitative genetics. *Trends Genet TIG* **29**: 669–76.

Nithianantharajah J, Hannan AJ. 2007. Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays News Rev Mol Cell Dev Biol* **29**: 525–35.

Nusinow D a., Helfer A, Hamilton EE, King JJ, Imaizumi T, Schultz TF, Farré EM, Kay S a. 2011. The ELF4–ELF3–LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*.

O'Dushlaine CT, Edwards RJ, Park SD, Shields DC. 2005a. Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol* **6**: R69.

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, et al. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**: 1619–1622.

Oka A, Hayashi H, Tomizawa M, Okamoto K, Suyun L, Hui J, Kulski JK, Beilby J, Tamiya G, Inoko H. 2003. Localization of a non-melanoma skin cancer susceptibility region within the major histocompatibility complex by association analysis using microsatellite markers. *Tissue Antigens* **61**: 203–210.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**: 92–94.

Parsons PA. 1992. Fluctuating asymmetry: a biological monitor of environmental and genomic stress. *Heredity* **68 (Pt 4)**: 361–364.

Peixoto a a, Hennessy JM, Townson I, Hasan G, Rosbash M, Costa R, Kyriacou CP. 1998a. Molecular coevolution within a Drosophila clock gene. *Proc Natl Acad Sci U S A* **95**: 4475–80.

Peixoto A, Campesan S, Costa R, Kyriacou C. 1993. Molecular evolution of a repetitive region within the per gene of Drosophila. *Mol Biol Evol* **10**: 127–139.

Persi E, Horn D. 2013. Systematic analysis of compositional order of proteins reveals new characteristics of biological functions and a universal correlate of macroevolution. ed. J.S. Fetrow. *PLoS Comput Biol* **9**: e1003346.

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, et al. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**: 368–372.

Poduri A, Evrony GD, Cai X, Walsh CA. 2013. Somatic mutation, genomic variation, and neurological disease. *Science* **341**: 1237758.

Poeta L, Fusco F, Drongitis D, Shoubridge C, Manganelli G, Filosa S, Paciolla M, Courtney M, Collombat P, Lioi MB, et al. 2013. A regulatory path associated with X-linked intellectual disability and epilepsy links KDM5C to the polyalanine expansions in ARX. *Am J Hum Genet* **92**: 114–25.

Ponder RG, Fonville NC, Rosenberg SM. 2005. A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation. *Mol Cell* **19**: 791–804.

Posadas DM, Carthew RW. 2014. MicroRNAs and their roles in developmental canalization. *Curr Opin Genet Dev* **27C**: 1–6.

Prentice AM, Gershwin ME, Schaible UE, Keusch GT, Victora CG, Gordon JI. 2008. New challenges in studying nutrition-disease interactions in the developing world. *J Clin Invest* **118**: 1322–1329.

Press M, Carlson KD, Queitsch C. 2014. *The overdue promise of short tandem repeat variation for heritability*. http://biorxiv.org/lookup/doi/10.1101/006387 (Accessed July 10, 2014).

Preston BD, Albertson TM, Herr AJ. 2010. DNA replication fidelity and cancer. *Semin Cancer Biol* **20**: 281–293.

Qi J, Zhao F. 2011a. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* **39**: W567–575.

Queitsch C, Carlson KD, Girirajan S. 2012a. Lessons from model organisms: phenotypic robustness and missing heritability in complex disease. *PLoS Genet* **8**: e1003041.

Queitsch C, Sangster TA, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**: 618–624.

Rando OJ, Verstrepen KJ. 2007. Timescales of genetic and epigenetic inheritance. *Cell* **128**: 655–668.

Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences, and control. *Science* **309**: 2010–2013.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–60.

Rival P, Press MO, Bale J, Grancharova T, Undurraga SF, Queitsch C. 2014. *The conserved PFT1 tandem repeat is crucial for proper flowering in Arabidopsis thaliana*. http://biorxiv.org/lookup/doi/10.1101/006437 (Accessed July 10, 2014).

Rohlfs RV, Fullerton SM, Weir BS. 2012. Familial identification: population structure and relationship distinguishability. *PLoS Genet* **8**: e1002469.

Rosas U, Mei Y, Xie Q, Banta JA, Zhou RW, Seufferheld G, Gerard S, Chou L, Bhambhra N, Parks JD, et al. 2014a. Variation in Arabidopsis flowering time associated with cis-regulatory variation in CONSTANS. *Nat Commun* **5**: 3651.

Rubinsztein DC, Leggo J, Chiano M, Dodge A, Norbury G, Rosser E, Craufurd D. 1997. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. *Proc Natl Acad Sci* **94**: 3872–3876.

Rutherford S, Hirate Y, Swalla BJ. 2007. The Hsp90 capacitor, developmental remodeling, and evolution: the robustness of gene networks and the curious evolvability of metamorphosis. *Crit Rev Biochem Mol Biol* **42**: 355–372.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. *Nature* **396**: 336–342.

Rutter MT, Roles A, Conner JK, Shaw RG, Shaw FH, Schneeberger K, Ossowski S, Weigel D, Fenster CB. 2012. Fitness of Arabidopsis thaliana mutation accumulation lines whose spontaneous mutations are known. *Evol Int J Org Evol* **66**: 2335–2339.

Salathia N, Queitsch C. 2007. Molecular mechanisms of canalization: Hsp90 and beyond. *J Biosci* **32**: 457–463.

Sangster TA, Bahrami A, Wilczek A, Watanabe E, Schellenberg K, McLellan C, Kelley A, Kong SW, Queitsch C, Lindquist S. 2007. Phenotypic diversity and altered environmental plasticity in Arabidopsis thaliana with reduced Hsp90 levels. *PloS One* **2**: e648.

Sangster TA, Lindquist S, Queitsch C. 2004. Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. *BioEssays News Rev Mol Cell Dev Biol* **26**: 348–362.

Sangster TA, Salathia N, Lee HN, Watanabe E, Schellenberg K, Morneau K, Wang H, Undurraga S, Queitsch C, Lindquist S. 2008a. HSP90-buffered genetic variation is common in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **105**: 2969–2974.

Sangster TA, Salathia N, Undurraga S, Milo R, Schellenberg K, Lindquist S, Queitsch C. 2008b. HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. *Proc Natl Acad Sci U S A* **105**: 2963–2968.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. ed. R.W. Sobol. *PloS One* **8**: e54710.

Sawyer LA. 1997. Natural Variation in a Drosophila Clock Gene and Temperature Compensation. *Science* **278**: 2117–2120.

Sawyer LA, Hennessy JM, Peixoto AA, Rosato E, Parkinson H, Costa R, Kyriacou CP. 1997. Natural variation in a Drosophila clock gene and temperature compensation. *Science* **278**: 2117–2120.

Scarpino SV, Hunt PJ, Garcia-De-Leon FJ, Juenger TE, Schartl M, Kirkpatrick M. 2013a. Evolution of a genetic incompatibility in the genus Xiphophorus. *Mol Biol Evol* **30**: 2302–2310.

Schaefer MH, Wanker EE, Andrade-Navarro MA. 2012. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Res* **40**: 4273–87.

Schaper E, Gascuel O, Anisimova M. 2014. Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Mol Biol Evol* **31**: 1132–1148.

Scheib JE, Gangestad SW, Thornhill R. 1999. Facial attractiveness, symmetry and cues of good genes. *Proc Biol Sci* **266**: 1913–1917.

Schmitt MW, Prindle MJ, Loeb LA. 2012. Implications of genetic heterogeneity in cancer. *Ann N Y Acad Sci* **1267**: 110–116.

Schuermann D, Fritsch O, Lucht JM, Hohn B. 2009. Replication Stress Leads to Genome Instabilities in Arabidopsis DNA Polymerase Mutants. *PLANT CELL ONLINE* **21**: 2700–2714.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316**: 445–449.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**: 1038–1042.

Shee C, Gibson JL, Darrow MC, Gonzalez C, Rosenberg SM. 2011. Impact of a stress-inducible switch to mutagenic repair of DNA breaks on mutation in Escherichia coli. *Proc Natl Acad Sci U S A* **108**: 13659–13664.

Slagboom PE, Mullaart E, Droog S, Vijg J. 1991. Somatic mutations and cellular aging: two-dimensional DNA typing of rat fibroblast clones. *Mutat Res* **256**: 311–321.

Smukalla S, Caldara M, Pochet N, Beauvais A, Guadagnini S, Yan C, Vinces MD, Jansen A, Prevost MC, Latgé J-P, et al. 2008. FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell* **135**: 726–37.

Sollars V, Lu X, Xiao L, Wang X, Garfinkel MD, Ruden DM. 2003. Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nat Genet* **33**: 70–74.

Specchia V, Piacentini L, Tritto P, Fanti L, D'Alessandro R, Palumbo G, Pimpinelli S, Bozzetti MP. 2010a. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature* **463**: 662–665.

Strachan DP. 2000a. Family size, infection and atopy: the first decade of the "hygiene hypothesis." *Thorax* **55 Suppl 1**: S2–10.

Strachan DP. 2000b. The role of environmental factors in asthma. *Br Med Bull* **56**: 865–882.

Subramanian S, Mishra R, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**: R13.

Sureshkumar S, Todesco M, Schneeberger K, Harilal R, Balasubramanian S, Weigel D. 2009. A genetic defect caused by a triplet repeat expansion in Arabidopsis thaliana. *Science* **323**: 1060–3.

Sutherland GR, Kremer E, Lynch M, Pritchard M, Yu S, Richards RI, Haan EA. 1991. Hereditary unstable DNA: a new explanation for some old genetic questions? *The Lancet* **338**: 289–292.

Tae H, McMahon KW, Settlage RE, Bavarva JH, Garner HR. 2013. ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats. *Bioinforma Oxf Engl* **29**: 1734–1741.

Taipale M, Jarosz DF, Lindquist S. 2010. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nat Rev Mol Cell Biol* **11**: 515–528.

Thornhill null, Gangestad null. 1999. Facial attractiveness. *Trends Cogn Sci* **3**: 452–460.

Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**: 213–223.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. 2009. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**: 315–316.

Undurraga SF, Press MO, Legendre M, Bujdoso N, Bale J, Wang H, Davis SJ, Verstrepen KJ, Queitsch C. 2012a. Background-dependent effects of polyglutamine variation in the Arabidopsis thaliana gene ELF3. *Proc Natl Acad Sci U S A* **109**: 19363–19367.

Undurraga SF, Press MO, Legendre M, Bujdoso N, Bale J, Wang H, Davis SJ, Verstrepen KJ,

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005a. Intragenic tandem repeats generate functional variability. *Nat Genet* **37**: 986–990.

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**: 1213–6.

Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. *BioEssays News Rev Mol Cell Dev Biol* **30**: 367–373.

Wagner A. 2000. Robustness against mutations in genetic networks of yeast. *Nat Genet* **24**: 355–361.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**: 539–543.

Whitacre JM. 2012. Biological robustness: paradigms, mechanisms, and systems principles. *Front Genet* **3**: 67.

Willems TF, Gymrek M, Highnam G, The 1000 Genomes Project, Mittelman D, Erlich Y. 2014b. *The Landscape of Human STR Variation*. http://biorxiv.org/lookup/doi/10.1101/004671 (Accessed July 15, 2014).

Yeyati PL, Bancewicz RM, Maule J, van Heyningen V. 2007. Hsp90 selectively modulates phenotype in vertebrate development. *PLoS Genet* **3**: e43.

Yu F, Sabeti PC, Hardenbol P, Fu Q, Fry B, Lu X, Ghose S, Vega R, Perez A, Pasternak S, et al. 2005a. Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS Genet* **1**: e41.

Yu J-W, Rubio V, Lee N-Y, Bai S, Lee S-Y, Kim S-S, Liu L, Zhang Y, Irigoyen ML, Sullivan J a, et al. 2008. COP1 and ELF3 control circadian function and photoperiodic flowering by regulating GI stability. *Mol Cell* **32**: 617–30.

Zhang Y, Liu C, Peng H, Zhang J, Feng Q. 2012. IL1 receptor antagonist gene IL1-RN variable number of tandem repeats polymorphism and cancer risk: a literature review and meta-analysis. ed. M. Katoh. *PloS One* **7**: e46017.

Zhao Z, Guo C, Sutharzan S, Li P, Echt CS, Zhang J, Liang C. 2014. Genome-wide analysis of tandem repeats in plants and green algae. *G3* **4**: 67–78.

Zühlke C, Dalski A, Schwinger E, Finckh U. 2005. Spinocerebellar ataxia type 17: report of a family with reduced penetrance of an unstable Gln49 TBP allele, haplotype analysis supporting a founder effect for unstable alleles and comparative analysis of SCA17 genotypes. *BMC Med Genet* **6**: 27.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**: 1193–8.