AN ABSTRACT OF THE THESIS OF

Anantnoor Brar for the degree of Honors Baccalaureate of Science in Biochemistry and Biophysics presented on August 12, 2013. Title: Dissecting DNA Damage Responses in Arabidopsis: A High-Throughput Sequencing Approach.

Abstract approved: _____
John B. Hays

A unique, permanent root-growth arrest phenotype was observed in UV-B irradiated *Arabidopsis thaliana* roots lacking the DNA damage response kinase ATR. Segregation analysis suggests the dependence of this phenotype on another, unidentified gene, termed *ursu*. High-throughput sequencing technologies were used to map and identify the causal *ursu* mutation. Illumina HiSeq 2000 short sequence reads were analyzed using bioinformatics command line tools for Next-Generation Mapping (NGM) and Mapping and Alignment with Short Sequences (MASS) programs. Several putative candidate genes were identified, including auxin response factor 13 (*arf13*) and resistance to *Peronospora parasitica* protein 8 (*rpp8*). Sanger sequencing verified the presence of a single nucleotide polymorphism (SNP) and a frameshift mutation caused by a 26-bp deletion in *arf13* in *ursu atr* plants. Although additional sequencing and complementation tests are still necessary, a causative mutation in *arf13* could provide clues linking the DNA damage response kinase ATR with the key growth and development phytohormone auxin.

Key Words: mapping-by-sequencing, ATR, root-development, UV-B, auxin
Corresponding e-mail address: brara@lifetime.oregonstate.edu

Dissecting DNA Damage Responses in Arabidopsis:

A High-Throughput Sequencing Approach

by

Anantnoor Brar

A PROJECT

submitted to

Oregon State University

University Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Biochemistry and Biophysics (Honors Associate)

Presented August 12, 2013
Commencement June 2014

Honors Baccalaureate of Science in Biochemistry and Biophysics project of Anantnoor Brar presented on August 12, 2013.

APPROVED:

_____

Mentor, representing Environmental and Molecular Toxicology

_____

Committee Member, representing Botany and Plant Pathology

_____

Committee Member, representing Environmental and Molecular Toxicology

_____

Chair, Department of Biochemistry and Biophysics

_____

Dean, University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, University Honors College. My signature below authorizes release of my project to any reader upon request.

_____

Anantnoor Brar, Author

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor John Hays, for his mentorship and for providing me the opportunity to work within the exciting field of DNA repair. Thank you for your advice throughout the past four years, and especially for your perseverance and strength during this past year – you are a source of inspiration. I would like to extend my appreciation to the members of my committee, Drs. Marc Curtis and Andrew Buermeyer, for their feedback and guidance. I would especially like to credit Dr. Curtis for involving me in this project and for his ideas, patience, and enthusiasm for science. Thank you, Drs. Gary Merrill and Chris Matthews, for your contributions in the final stages of my thesis examination. To past and present members of the Hays lab, thank you for your companionship and continued friendship. Particularly, I extend my appreciation to Peter Hoffman and Robert Ursu, for sharing their expertise. I would like to thank Nathan Lai, Drs. Thomas Wolpert and Jennifer Lorang for assistance with bioinformatics and thoughtful discussions. To Erin Bredeweg and Dr. Michael Freitag, thank you for your help with sequencing library preparation and funding. I am grateful for the outstanding bioinformatics support and resources available through the Center for Genome Research and Biocomputing, without which this project would not be possible. I am thankful for Dr. Shawn O'Neil's pedagogical approach to computer programming instruction. To my advisors Dr. Kevin Ahern, Dr. Indira Rajagopal, Claire Colvin, and Rebekah Lancelin, thank you for the dinners, lively discussions, and unconditional support and guidance. To my peers in the University Honors College and Biochemistry and Biophysics who have been a continual source of inspiration and have given me a standard of excellence to strive for in all endeavors, I am appreciative. I am indebted to Dr. Buck Wilcox for his

unwavering support throughout this project and in the preparation of this document, and for challenging me to be mindful, think critically, and to always be curious. Thank you, Arlyn, Ayeza, Yanet, and all of my friends, for your camaraderie and encouragement. Lastly, I would like to express gratitude for the endless patience, support, and love from my parents, siblings, and extended family throughout my work on this project.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABBREVIATIONS

9-1-1	RAD9-RAD1-HUS1 complex
ARF	Auxin response factor
ATM	Ataxia-telangiectasia mutated
ATR	ATM and RAD3-related
ATRIP	ATR-interacting protein
Aux	Auxin
BAM	Binary Alignment/Map format
BWA	Burrows-Wheeler Aligner (program)
CASHX	Cache Assisted Hash Search with Xor logic (program)
CDS	Coding DNA Sequence
ChD	Discordant chastity
Chk1	Checkpoint kinase 1
Col-0	Columbia-0 *Arabidopsis thaliana* ecotype
CPD	Cyclobutane pyrimidine dimer
DDR	DNA damage response
DSB	Double-strand break
EMS	Ethyl methanesulfonate
G1	Gap 1 phase (cell cycle)
G2	Gap 2 phase (cell cycle)
HNPCC	Human non-polyposis colorectal cancer
IAA	Indole-3-acetic acid
IR	Ionizing radiation
Ler	Landsberg *erecta Arabidopsis thaliana* ecotype

LIST OF ABBREVIATIONS (Continued)

M	Mitosis phase (cell cycle)
MAQ	Mapping and Assembly with Quality (program)
MASS	Mapping and Alignment with Short Sequences
MMR	Mismatch repair
NER	Nucleotide excision repair
NGM	Next-generation mapping
NIKS	Needle in the k-stack
PCD	Programmed cell death
PCR	Polymerase chain reaction
PIKK	Phosphoinositide 3-kinase (PI3K) related protein kinase
PIN	PIN FORMED
PLT	PLETHORA
QC	Quiescent center
QTL	Quantitative trait loci
RAM	Root apical meristem
RPA	Replication protein A
RPP8	Resistance to *Peronospora parasitica* protein 8
S	Synthesis phase (cell cycle)
SAM	Shoot apical meristem
SNP	Single nucleotide polymorphism
SOG1	Suppressor of gamma response 1
SSB	Single-strand break
StPr	Stem and progenitor cells
TA	Transiently amplifying cells

TAIR	The Arabidopsis Information Resource
T-DNA	Transfer-DNA
TOPBP1	Topoisomerase-binding protein-1
UV	Ultraviolet

ਇਹ ਖੋਜ ਪੇਪਰ ਮੈਂ ਆਪਣੇ ਸਤਿਕਾਰ ਯੋਗ ਦਾਦਾ ਜੀ, ਦਾਦੀ ਜੀ, ਨਾਨਾ ਜੀ ਅਤੇ ਨਾਨੀ ਜੀ ਨੂੰ ਅਰਪਣ ਕਰਦੀ ਹਾਂ। ਉਹਨਾਂ ਦੇ ਅਸ਼ੀਰਵਾਦ ਅਤੇ ਪਰੇਰਣਾ ਨੇ ਮੈਨੂੰ ਉੱਚ ਵਿਦਿਆ ਦੀ ਕਾਮਯਾਬੀ ਦਾ ਰਾਹ ਦਿਖਾਇਆ। ਮੈਂ ਉਹਨਾਂ ਦਾ ਤਹਿ ਦਿਲੋਂ ਧੰਨਵਾਦ ਕਰਦੀ ਹਾਂ।

For my beloved grandparents, who have always supported me in my academic endeavors and who taught me to put forth my best efforts, this thesis is dedicated to you.

Dissecting DNA Damage Responses in Arabidopsis: A High-Throughput Sequencing Approach

Introduction

1.1 Mutation and the DNA Damage Response

Cells are continuously bombarded with mutagens, both endogenous and exogenous in origin. Endogenous sources include reactive oxygen species (as by-products of aerobic respiration and heavy metal oxidation during Fenton reactions), DNA polymerase replication errors (10^{-6} – 10^{-7} per base pair replicated), spontaneous deamination and depurination of DNA bases, and DNA strand breaks caused by abortive topoisomerases [1, 2]. Exogenous mutagens include UV- and ionizing-radiation, aflatoxins, and polycyclic aromatic hydrocarbons (e.g. compounds found in diesel exhaust and cigarette smoke). As a result, tens of thousands of pre-mutagenic DNA lesions are created every day. Lesions can pose an impediment to DNA transcription and replication that may lead to further mutation or cell death if the cause of stalled nucleotide synthesis is not removed or bypassed by DNA damage response mechanisms [1]. Damaged DNA can lead to DNA single-strand breaks (SSBs) and DNA double-strand breaks (DSBs). Importantly, accumulated DNA lesions become fixed as mutations, which antagonize genome fidelity and may increase risk of cancer development. DNA damage impedes the accurate passage of genetic information from progenitor to offspring. DNA repair and damage tolerance mechanisms, collectively called the DNA-damage response (DDR), exist within cells to prevent the accumulation of potentially hazardous DNA lesions.

To maintain genome stability, DNA repair systems must be able to handle diverse types of damage. This is done through several, distinct repair pathways. Mismatch repair (MMR) corrects base-base mismatches and insertion/deletion loop-outs that are created during DNA replication and remain uncorrected after the polymerase proofreading [3]. MMR also antagonizes homeologous recombination, which, if left unchecked, results in gene conversion or non-reciprocal crossover. Nucleotide excision repair (NER) repairs bulky or helix-distorting base lesions, such as cyclobutane pyrimidine dimers (CPDs) and pyrimidine [6-4] pyrimidinone dimers induced by UV exposure, as well as DNA cross-links and certain kinds of chemically damaged bases [4, 5]. Base excision repair complements NER by correcting smaller, non-helix-distorting lesions and repairing abasic sites [6]. Occasionally, DNA damage is tolerated and bypassed during DNA replication by error-prone translesion polymerases [1]. In the presence of a sister chromatid to guide repair, such as during synthesis (S) phase and mitosis, homologous recombination is the favored repair pathway for the repair of double-strand DNA breaks [7]. If a sister chromatid is not available, (e.g. during interphase Gap 1 of the cell cycle before DNA synthesis initiates in S phase), non-homologous end-joining is used in the repair of DSBs [8]. The Fanconi anemia pathway repairs inter-strand DNA cross-links [9]. Kinases ataxia-telangiectasia mutated (ATM) and ATM and RAD3-related (ATR) are the major regulators of the DDR. ATM and ATR modulate downstream cellular responses through phosphorylation, sumoylation, ubiquitylation, and acetylation of substrates to coordinate cell cycle transitions, DNA repair, and apoptosis [10].

Cells deficient in DNA repair mechanisms are especially susceptible to the accumulation of mutations. Although some mutations are neutral (being neither detrimental nor

beneficial) and some are beneficial to organismal fitness, the vast majority of mutations are deleterious [11]. The diploid rate of genomic mutations affecting organismal fitness in *Arabidopsis thaliana* is estimated at 0.2 ± 0.1 per generation [12]. Of the mutations affecting fitness, ~70% are estimated to be deleterious (0.14 ± 0.04 per generation) in Arabidopsis. Mutations in DNA repair genes are generally detrimental to an organism. Aberrant MMR, for instance, is associated with human non-polyposis colorectal cancer (HNPCC) [13]. Other genes whose deficiencies cause organisms or cells to accumulate mutations at an increased rate are said to be mutators. For example, mutant alleles of the exonucleolytic (proofreading) subunit of DNA polymerase, mutT and its homologs, and replicative DNA helicase are known to cause a mutator phenotype [14, 15].

An improved understanding of DNA damage responses facilitates the development of medical diagnosis and treatment. DNA repair machinery safeguards genetic information against the damage inflicted upon DNA by mutagens. Radiation, aflatoxins, and polycyclic aromatic hydrocarbons cause DSBs and various cancers (most notably those of the lungs and oral cavity) if left unrepaired [1]. Between two and seven percent of all hereditary colorectal cancers are associated with defective MMR [13]. Inherited deficiencies in MMR genes cause the most common form of hereditary colorectal cancer, HNPCC. Early detection of mutations that predispose individuals to genetic diseases and cancers may reduce the risk of morbidity or mortality through preventative, therapeutic, or lifestyle interventions.

Accurate plant propagation, where mutation is the bane of horticulturists trying to maintain stable varieties of plants, may also be informed by understanding DNA repair mechanisms. In contrast, agricultural trait discovery and development uniquely benefit

from mutagenesis. Generally, mutation is viewed negatively because it causes disease and destabilizes valuable germplasm. However, plant and animal breeders rely on mutagenesis to provide the raw material on which they may select and improve novel traits. One specific area in plant breeding that may benefit from a better understanding of DDR is induction of new mutations as is currently done with radiation and chemical mutagens. Currently, plant breeders use chemical mutagens such as ethyl methanesulfonate (EMS) to induce novel mutations in crops [16]. EMS mutagens preferentially alkylate guanine residues, forming O^6-ethylguanine. This product tends to pair with thymine rather than cytosine, thus leading to non-canonical base pairing and C/G → T/A substitutions. EMS is called a "supermutagen," a term coined by Rapoport et al. [17] for highly mutagenic chemicals with low toxicity that do not cause much chromosome breakage. EMS mutagenesis is not gene-specific – it acts throughout the genome. Plants containing dysfunctional MMR or defective replicative DNA polymerases may accumulate mutations at an increased rate and expose the genome to a spectrum of mutations broader than can be obtained by chemical mutagenesis [18]. In the future, especially if treatments that produce mutators are found to induce substantive numbers of novel mutations other than C/G → T/A substitutions, these techniques could change breeding practices. Breeders may be able to use a defective mismatch repair system to induce and select for novel traits in their crops and livestock, and then subsequently restore MMR function, resulting in an individual free of recombinant DNA.

1.2 ATR

Phosphoinositide 3-kinase (PI3K) related protein kinases (PIKKs) regulate the DNA damage response in cells [19]. Sometimes referred to as "sentries at the gate of genome

stability," ataxia-telangiectasia mutated (ATM) kinase, and ATM and RAD3-related (ATR) kinase promote signaling, cell-cycle arrest, and DNA repair in response to DNA damage [20]. ATM and ATR are crucial to the proper functioning of DNA damage cell cycle checkpoints, including regulating the G1-S and G2-M transitions. In response to DNA damage such as that induced by ultraviolet (UV) or ionizing radiation (IR), activation of checkpoints temporarily pauses the cell cycle to allow sufficient time for cells to repair their damaged DNA, bypass the damage, or undergo apoptosis [21]. ATM and ATR kinases share significant sequence homology and target an overlapping set of substrates to repair DSBs and rectify stalled DNA replication. Double-strand breaks primarily activate ATM functionality. In contrast, ATR is activated during every S phase of the cell cycle to promote replication-fork stability, apoptosis, or prevent entry into mitosis in the presence of DNA lesions or stalled replication forks. A variety of DNA damaging agents can activate ATR, including UV light, alkylating agents, and chemical inhibitors of DNA replication [22].

ATR deficiency is lethal in early embryonic stages in mammals, whereas ATM deficiency can go unnoticed for decades in many individuals. Mutations in ATM are found in 0.5% – 1% of the population and several of these mutations may lead to the development of cancers. Ataxia-telangiectasia, a pathological condition found in individuals with homozygous mutations in ATM, is a rare neurodegenerative disorder that causes severe mental retardation and disturbances in coordination [23]. As ATR mutations are almost always lethal in mammals, viable mutants are rare and only occur in heterozygous carriers or as hypomorphic mutations. Seckel syndrome, an exceedingly rare condition (affecting approximately one in 10,000 live births), is one of the few links

between mutations in ATR and disease, and it is characterized by severe growth and mental retardation, short stature, and microcephaly [24].

ATR plays a crucial role in the maintenance of genome fidelity. As noted above, ATR protein kinase is activated by a wide variety of DNA lesions, as well as by stalled replication forks. Once activated, ATR and ATM phosphorylate an overlapping set of downstream proteins. ATR-specific targets include p53 and checkpoint kinase 1 (Chk1). Phosphorylation of these targets eventually blocks critical cell-cycle transitions, providing the cell with a crucial window of time in which to respond to DNA damage.

ATR is activated by persistent single-stranded DNA (ssDNA) coated by Replication Protein A (RPA) [25]. RPA-coated ssDNA formed during DNA replication and DNA repair is important in localizing ATR to the site of damage. However, RPA-ssDNA and ATR are not the only participants in this scheme. ATR-interacting protein (ATRIP) binds to RPA and is considered so crucial for ATR recognition of RPA-ssDNA that it is essentially treated as a subunit of ATR. ATR-ATRIP activation and signaling are further dependent on colocalization with the RAD9-RAD1-HUS1 (9-1-1) complex. The 9-1-1 complex, similar in structure and function to the proliferating cell nuclear antigen (PCNA) complex involved in DNA replication, is recruited to the 5’ junction of double-stranded DNA (dsDNA) adjacent to RPA-ssDNA. In order to stimulate ATR signaling, the 9-1-1 complex recruits topoisomerase-binding protein-1 (TOPBP1) to the ATR complex in a RAD9 C-terminal phosphorylation-dependent manner. TOPBP1 contains an ATR activation domain which has been shown to activate ATR-ATRIP complexes *in vitro*, although the exact mechanism of activation is unknown [26]. Some evidence exists

that lesions are recognized by ATR directly. In a partially reconstituted system, ATR has been demonstrated to bind to and become activated by UV-damaged DNA [27].

1.3 *Arabidopsis thaliana*

The advantages of *Arabidopsis thaliana* as a model system to study UV mutagenesis, DDR, cell signaling and translesion synthesis are manifold. Arabidopsis, a member of the *Brassicaceae* family, is a small, diploid angiosperm [28]. Although Arabidopsis does not hold significant agronomic value, it has been widely studied as a model organism in plant biology. Several characteristics of Arabidopsis enable efficient, classical genetic analysis in the plant. Arabidopsis is a small plant with minimal growth requirements, allowing many plants to be grown in a small space under fluorescent illumination, and it also grows well on agar growth medium in Petri plates. The plant is a prolific seed producer, making up to 20,000 seeds on a single plant about six weeks after germination. The ease of obtaining up to a few hundred seeds by manually crossing individuals, efficient transformation techniques using *Agrobacterium tumefaciens*, large numbers of mutant lines and genomic resources available, and the small, sequenced and heavily annotated genome (125 Mb haploid) greatly facilitate genetic analysis [29]. *Agrobacterium tumefaciens* is a plant pathogen that induces crown gall tumors by injecting a plasmid, called tumor inducing principle, into plant cells [30]. Agrobacterium contains a large Ti plasmid, and tumor induction is a direct consequence of T-DNA incorporation into the plant genome. A valuable genetic tool, Agrobacterium Ti plasmids have been disarmed of their oncogenes and are commonly transformed into Arabidopsis in order to disrupt existing genes or introduce novel genes. Despite its small genome, the structure of individual Arabidopsis genes and chromosomes, genetic properties, and the overall

complement of genes in the Arabidopsis genome are typical of those of other flowering plants. Additionally, there is a high level of polymorphism between the various accessions of Arabidopsis, which facilitates mapping experiments. Heavily studied accessions, or ecotypes, include Columbia, Landsberg *erecta*, and Wassilewskija. Paramount to this study, ATR deletion in Arabidopsis does not lead to embryonic lethality, unlike the situation in animals. ATR-deficient plants may be able to survive because their developmental program is simpler and more plastic than that of animal development. In other words, plants do not have reduced reliance on genes for genome maintenance, rather, they are better able to tolerate the effects of missing genes [31]. As with other model organisms, the interest is not specific to Arabidopsis, but rather, in what the model can reveal about basic biology and, in this study, the DNA damage response.

1.4 Root Apical Meristem

The root meristem has emerged as a model to study stem cells and the stem-cell niche. Like most angiosperms, Arabidopsis produces an extensive root system designed to anchor the plant and absorb nutrients. Like their above ground counterparts in the shoot apical meristem (SAM), which contains a stem cell niche, roots also possess a stem cell niche at the growing tip in the root apical meristem (RAM) [32]. Along the apical-basal axis of the root, different zones are responsible for cell division, cell expansion, and cell differentiation. The RAM is located at the apex of the root and houses the stem-cell niche of continuously dividing cells which produce the basic cell types and define their organization.

Although plant roots are not typically exposed to UV-B, roots are a convenient system of study for many reasons: the root apical meristem is physically accessible (i.e. the

meristem is not shielded by developing organs or primordia as is the SAM), the root is free of pigments and therefore essentially transparent, there are relatively few differentiated cell types in roots, and the orderly arrangement of cells in vertical "files" in roots facilitates tracing the origin of each individual cell back to its origin in the quiescent center (QC) of the RAM [33]. The hierarchical organization of the Arabidopsis stem cell niche is characteristic of both plant and animal tissues and closely resembles the mammalian colonic crypt [34]. Stem cells, also called initials, surround one or two quiescent center cells (Figure 1). Cells of the QC divide very infrequently and are a source of replenishment for the more rapidly cycling initials. In a single, asymmetric stem cell division, one daughter remains a stem cell, while one daughter becomes a progenitor cell [35]. Together, the stem and progenitor cells (collectively referred to as StPr) and the QC constitute the stem cell niche. Progenitor cells continue dividing to become transiently amplifying (TA) cells, which in turn undergo repeated divisions before being displaced away from the growing root tip into the root's elongation zone. Controlled cell expansion in the root elongation zone results in the characteristic elongated shape of root cells and also provides the force that drives the root apex forward. At the top of the meristem-niche hierarchy, and thus progenitors of all downstream cells, stem cells must strictly maintain genome fidelity in the face of DNA lesions. Complications hindering previous study of the DDR in stem cell niches in animals highlight the advantages of using a simple plant system that is viable even in the absence of DDR genes essential in animals.

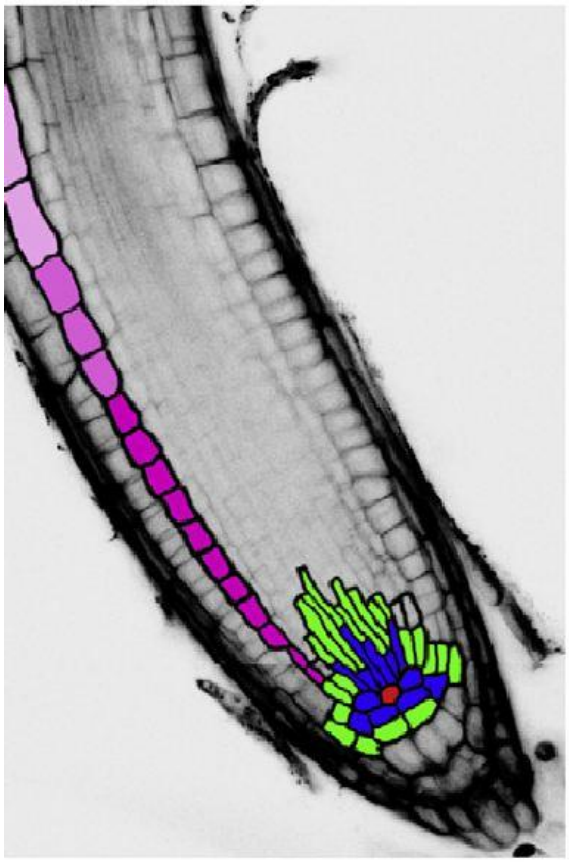


Figure 1. Arabidopsis root stem cell niche. Stem cells (blue) and their immediate daughters, the progenitor cells (green), divide to give rise to transiently amplifying cells (dark pink) which stop dividing and begin to mature (lighter pinks) in response to hormone gradients. The quiescent center is populated by one or two cells (red) [36].

1.5 Ultraviolet-B

Although the ozone layer of the stratosphere effectively blocks almost all UV-C (<280 nm), the radiation most efficiently absorbed by DNA, and blocks some UV-B (280 – 320 nm), the levels of UV radiation are not reduced to nonlethal levels [37]. Thus, both animals and plants maintain defense mechanisms to protect themselves from UV-A (320 – 400 nm) and UV-B: physical shielding by phenylpropanoids and flavonoids, direct reversal by photolyase, and DDR pathways designed to remove damage caused by the small amount of radiation that manages to reach the nucleus. The absence of particular DDR systems, such as NER defects in the disease Xeroderma pigmentosum, can make UV irradiation from the sun lethal in placental mammals [38]. Sessile plants face significant challenges in dealing with UV irradiation, and have developed unique mechanisms to maintain genomic integrity: physically shielding the shoot apical meristem, shedding photosynthetic tissues on an annual basis, producing UV-protective pigments such as flavonoids, and expressing DDR systems specialized to deal with UV-induced DNA lesions [37]. UV radiation is a threat to genome stability primarily because it creates pyrimidine dimers in DNA, the most prevalent of which are cyclobutane pyrimidine dimers (CPDs), which account for approximately 75% of UV-damaged bases [37, 39]. Pyrimidine [6-4] pyrimidinone dimers constitute the remainder. Both pyrimidine dimers block DNA replication and transcription, and if ssDNA persists, DSBs may result. Although such dimers can be tolerated and bypassed by translesion polymerases to allow DNA replication to continue, accuracy is often compromised and point mutations result. Mutagenesis is a serious result of UV-induced DNA damage, but the effects of pyrimidine dimers may be more detrimental to the survival of the organism.

Nonreplicating, terminally differentiated cells (such as root hair cells and neurons) need to remove dimers from the path of an oncoming RNA polymerase [37]. Thus, in the pursuit of genomic fidelity and to resume nucleic acid synthesis, it is essential for cells to effect the removal of premutagenic pyrimidine-dimer lesions.

1.6 Discovery of *ursu*

DNA lesions can have serious downstream consequences, including blockage of DNA replication that disturbs the timely progression of cell division in transiently amplifying cells and stem cells. In the somatic tissues of multicellular eukaryotes, TA cells and stem cells are largely responsible for tissue growth and regeneration. When cell division is interrupted by DNA damage, DNA damage response systems act to resume replication through removal of lesions or filling in gaps when replication reinitiates downstream of lesions. Cell-cycle delay and arrest, or cell death can result from unfilled gaps, DSBs, or stalled replication forks. To study the effects of DNA-damaging agents in proliferating-differentiating tissues in multicellular organisms, Furukawa et al. [40] induced programmed cell death (PCD) in the root stem-cell niche of *Arabidopsis thaliana* using UV-B and ionizing radiation (IR). The authors hypothesized that UV-B and IR would elicit root stem-cell death, and that PCD would be the result of a pathway requiring the transcription factor Suppressor of Gamma Response 1 (SOG1), ATR, or ATM. PIKKs were found to be a requirement for UV-B- or IR-induced programmed cell death; either ATR or ATM was sufficient to initiate PCD. A UV-B dose of 0.6 kJ m^{-2} elicited elevated stem-cell death in plants lacking ATR as compared to equally irradiated wild-type root tips. The increase in cell death per root in the absence of ATR may make sense considering the roles of ATR as a cell-cycle checkpoint regulator and a stabilizer of

collapsed and blocked DNA replication forks. Programmed cell death is an important protective mechanism that restores tissue homeostasis in stem cell niches and prevents the accumulation of irreparably damaged cells in tissues, albeit accompanied by a delay in growth. Both ATM and ATR can signal genome repair (in response to blocked replication forks or DSBs) or PCD in response to replication stress. Persistent, aberrant DNA structures initiate signaling of PCD. SOG1, a transcription factor previously shown to play an important role in the response to DNA-damaging agents such as UV-B and IR in Arabidopsis, was found to be required for both UV-B- and IR-induced PCD [41]. Despite significant differences in primary lesions, various DDR systems appear to converge at SOG1, ATM, or ATR to signal the fate of DNA-damaged cells.

These findings provided clues about the cooperation between DNA-damage activated protein kinases and other proteins they may interact with, such as SOG1. Combining two or more mutations in one organism (i.e. double and triple mutants) is a classic genetics approach used to determine activities and relative positions of proteins in pathways. Building on findings of involvement of ATR, ATM, and SOG1 in UV-B- and IR-induced PCD, and in an effort to elucidate the mechanisms of DNA damage signaling and protein activities in the relevant pathways, double mutants lacking ATR and SOG1 were pursued. In the process of creating the *atr sog1* double mutant, a novel mutant was discovered – one that in the absence of ATR and the presence of SOG1, caused root tip growth-arrest after UV-irradiation (Figure 2).

1.7 Genetic Mapping

Mapping genes to a particular locus is performed using either morphological (classical) or molecular markers. Genetic mapping is based on estimating the average number of

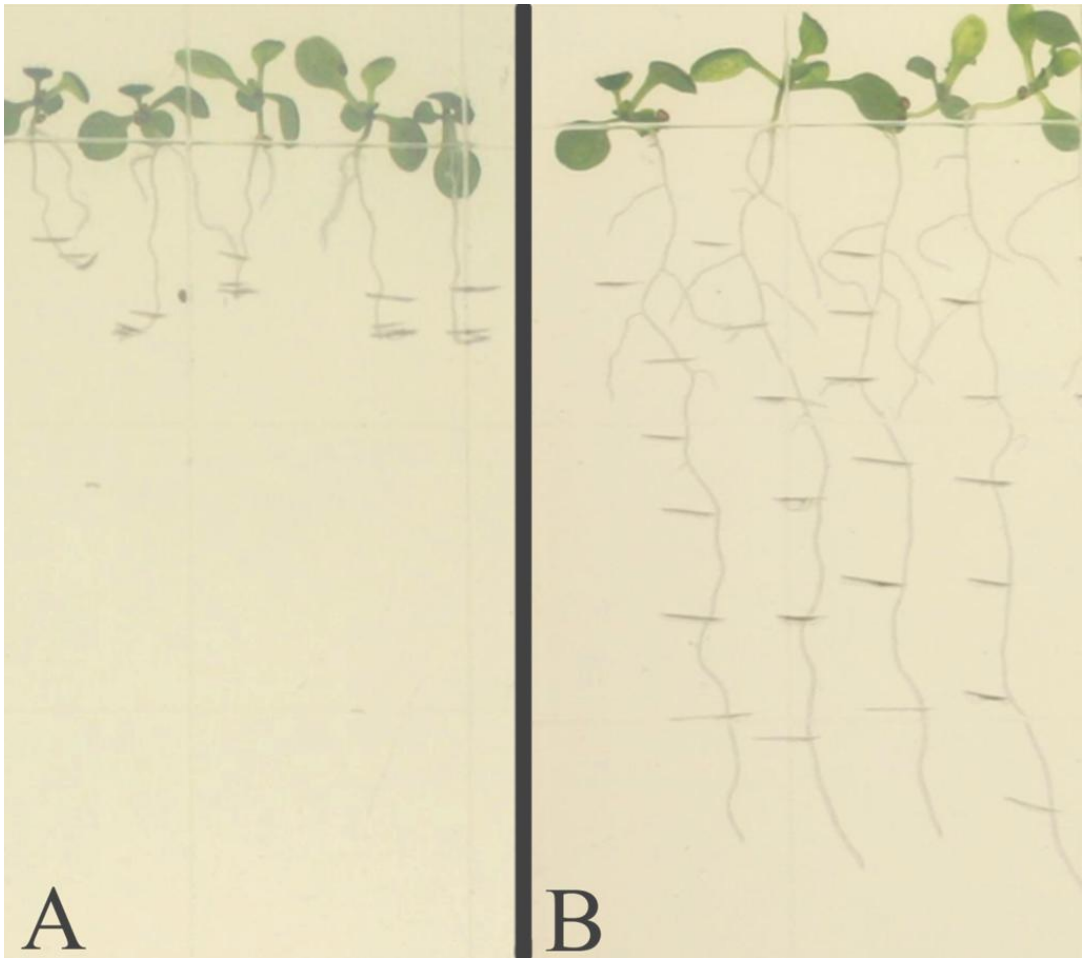


Figure 2. Root termination phenotype. Mutant *ursu atr* (A) and wild type (B) Arabidopsis thaliana root tips irradiated with 0.03 kJ m^{-2} UV-B. Vertical root growth was tracked every 24 hours, beginning three days after sowing and continuing through the UV-B irradiation and recovery phases, by marking the position of the root tip on the back of the Petri plate (Ursu R., unpublished). Mutant *atr* root tips show an initial delay in growth following UV-B irradiation, but they soon recover and continue growth at a similar rate to wild-type plants subjected to identical conditions (Curtis M., unpublished).

crossovers between two marker loci. Crossover events resulting in recombination between markers will separate the traits or markers from each other; separation of tightly linked traits is improbable. The map position of a gene of interest is determined using estimates of recombination distances for multiple pairs of markers [42]. Whether using morphological or molecular approaches, test-cross populations are used because each individual can easily be scored as recombinant or parental. Commonly in Arabidopsis, a new trait is mapped by crossing the homozygous recessive F2 mutant with a mapping line containing previously localized molecular marker genes. The polymorphisms contained in the mapping line can be morphological or molecular. The frequency that these markers are associated with the new trait provides a rough estimate of the chromosomal location of the causative mutation, given by the frequency of recombination. The frequency of recombination ranges from zero percent – indicating the two genes are completely linked (there is no recombination between the genes) – to fifty percent – which indicates that the two genes undergo independent assortment (i.e. they are on different chromosomes, or on opposite ends of the same chromosome). Recombination frequency is affected by the physical distance between the genes, as well as by the frequency of crossover events in the physical space between the genes. Refinement of the location to a specific locus is done by crossing mutant plants with individuals containing traits or markers near the locus identified in the initial mapping experiments [42].

The use of classical, or morphological, markers for mapping experiments does not require the use of technical molecular biology, and the experiments are, in principle, simple to execute. However, scoring phenotypes used for classical markers (e.g. flower color, plant stature, etc.) can be ambiguous and interference may occur between the marker

phenotype and the new trait being mapped. In addition, the number of markers that can be reliably scored in the progeny of a single cross is limited. Molecular markers include nucleotide sequence variation, such as restriction fragment length polymorphisms and microsatellite length polymorphisms, and their use in mapping offers a number of advantages compared to classical markers. Molecular markers are less likely to interfere with the mutant phenotype, and there are many molecular markers distributed across the genome. Thus, gene mapping is rapid and accurate. Linkage analysis using molecular markers involves crossing one ecotypic line showing the mutant trait with a different ecotypic line lacking the trait but encoding numerous polymorphic loci.

Although both classical and molecular markers have been used successfully to construct linkage maps for Arabidopsis, linkage mapping is laborious and time-intensive. On average, mapping a gene using either technique can take months to years, and it may take much longer, depending on the area being mapped [43]. In the wake of recent developments in high-throughput DNA sequencing technologies, new mapping techniques are beginning to emerge. As with mapping using molecular markers, most mapping by sequencing approaches require crossing the line containing the mutant of interest with a different ecotype, in order to use the naturally occurring polymorphisms between the two accessions to map the trait. The genomic DNA from F2 progeny of the mapping cross is sequenced, and the sequencing reads are aligned to a reference genome and analyzed using bioinformatics software. Advances in DNA sequencing technologies have greatly facilitated forward genetic mapping; identification of causal mutations using high-throughput sequencing and bioinformatics potentially decreases the time it takes to map a gene from years to months!

1.8 Sequencing Technologies and Genetic Mapping by Sequencing

The rapid advancement of DNA sequencing technologies over the past several decades has revolutionized the field of genetics. The advent of genomic DNA sequencing began in 1977, with the 5,386-bp bacteriophage φX174 genome making history as the first genome to be sequenced entirely [44]. At the time, DNA sequencing was a tedious, manual process performed using chain-terminating dideoxynucleotides. Although Sanger sequencing methods gradually became more automated with the development of capillary electrophoresis, the critical turning point for DNA sequencing required deviation from established technologies. These novel technologies, termed high-throughput sequencing and/or next-generation sequencing, are designed to sequence thousands, even millions, of DNA molecules in parallel [45]. Beginning in 2005, a handful of next-generation sequencing platforms became available, each with different sequencing chemistry. Such platforms included the Roche/454 GS FLX Titanium sequencer, the Illumina Genome Analyzer II/IIx, Applied Biosystems SOLiD, and Helicos HeliScope [44]. Differences in platform chemistry are suitable for solving different problems. Overall, the time to complete sequencing runs, regardless of next-generation sequencing platform, has been drastically reduced from several hours to Sanger sequence <1000 bp to several days to sequence millions of base pairs [46]. As sequencing technologies continue to evolve, sequencing costs become increasingly affordable; as of 2012, the cost per million bases using the Illumina HiSeq 2000 was $0.07.

The Illumina HiSeq 2000, used in this experiment, produces the industry's highest sequencing output: 600 Gb per run [46]. The HiSeq 2000 uses the same basic sequencing chemistry as its predecessors – reversible terminator-based sequencing by synthesis [47].

Sequencing by synthesis utilizes fluorescently labeled dNTPs to sequence millions of dsDNA clusters in parallel on the flow cell. First, single stranded PCR-amplified genomic DNA fragments ligated to adapter DNA molecules are adhered to the flow cell channel, which is densely populated with primers. Addition of unlabeled nucleotides and enzyme initiates bridge amplification of DNA, which is followed by denaturation of dsDNA fragments. Multiple rounds of amplification are performed to generate millions of clusters containing several million ssDNA fragments in each flow cell channel. Sequencing cycles require the addition of fluorescently labeled reversible terminators (dNTPs), DNA polymerase, and primers. Upon incorporation of fluorescently labeled dNTPs, the first base can be identified by the fluorescence emission following laser excitation. Each subsequent sequencing cycle follows this same format, sequencing each base in the fragment one at a time. Once sequencing of the fragments is complete, the short sequence reads in FASTQ format can be used for bioinformatics analysis, which may include assembly and alignment to a reference.

Powerful next-generation sequencing technologies with the ability to generate vast amounts of data have necessitated progress in the development of tools used for their analysis. Bioinformatics, an interdisciplinary field spanning computer science and biology, lies at the crux of sequencing technologies. As aforementioned, sequencing platforms have the ability to output hundreds of gigabytes of sequence information. Such unfathomably large datasets have become the trend in modern biology and it is no longer practical to manually analyze these datasets. Bioinformatics, then, becomes the tool of choice for the analysis of the gargantuan data output generated by sequencing platforms.

Advances in DNA and RNA sequencing have spurred the creation of entirely new fields of biology. Next-generation sequencing technologies have enabled mapping of epigenetic marks, accurate measurement of genome-wide transcript levels, an improved understanding of the structural organization of genomes, and helped identify sites of protein-DNA interactions [45]. Increased affordability and throughput of sequencing technologies are paving the path toward personalized medicine in the near future, which will likely involve the creation and analysis of individualized "omics" profiles to discover prominent risk factors, prevention methods, and treatment options. Sequencing technology continues to hold powerful implications for cancer genomics, forensic genomics, microbial genomics, and agricultural genomics.

Genetic mapping techniques, which traditionally relied on morphological and/or molecular markers, have been transformed by recent advances in sequencing technology that allow researchers to acquire large amounts of genomic sequence data inexpensively. Mutations responsible for phenotypes of interest can now be directly identified by high-throughput sequencing, thus by-passing months of work required by traditional methods. So-called mapping by sequencing can be performed using a selection of analysis tools. Among those available are Mapping and Alignment with Short Sequences (MASS) and Next-Generation Mapping (NGM), each which have been used successfully to map and identify causative mutations [48, 49]. Other mapping tools have been developed and new mapping programs continue to be developed – each requiring less *a priori* information about the genomes.

The Mapping and Alignment with Short Sequences (MASS) pipeline was developed for mapping and identification of causal mutations in forward genetics screens using

Arabidopsis thaliana. MASS requires a known list of polymorphisms between two ecotypes: the ecotype in which the mutant was isolated and a mapping line. Use of MASS requires the mutant line to be crossed with a mapping line of a different ecotype, and homozygous F2 progeny (~50-100) displaying the mutant phenotype are selected for sequencing and analysis. MASS uses SOAP, CASHX, and MAQ software packages in its analysis [50-52]. Together, these programs compare sequence reads from pooled mutant plant DNA to reference "clusters" generated from roughly 200 bp of TAIR10 genomic Arabidopsis reference sequence flanking each of the 305,002 single nucleotide polymorphisms known between mutant (Col-0) and mapping (Ler) ecotypes. After the mutant sequence is aligned to these clusters of sequence with Col-0-Ler polymorphisms, counts of ecotype specific perfect alignments to each polymorphic cluster are used to create plots of Col-0 specific SNP enrichment. Further analysis is performed on areas enriched for the ecotype in which the mutant was isolated (indicated by a more positive peak). To identify causative mutations, all sequence reads are re-mapped to a manually determined region (not exceeding 2 Mb) of the genome, based on SNP enrichment plots.

Next-generation mapping (NGM) is a *de novo* mapping tool, requiring no previous list of polymorphisms between two ecotypes (mutant and mapping lines). Like MASS, NGM also requires a cross between the mapping line (of a different ecotype than in which the mutant was identified) and the mutant line. Sequence reads generated from the F2 mutants are directly aligned to a reference genome, which is the same as the ecotype in which the mutation was originally isolated. NGM is a web-based Java tool which generates plots of SNP frequencies based on how well mutant sequence aligns to the reference. If a base matches the reference base, and is supported by most or all sequence

reads covering it, the position is given a discordant chastity score of zero. A discordant chastity score of ~0.5, in which half of the reported sequences at a position match the reference base and half differ, is expected at ecotype specific SNPs as a result of recombination events between the mutant and mapping ecotypes. A discordant chastity score of 1.0 is given at a position where the reported base unanimously differs from the reference base. NGM software scans discordant chastity scores for regions deficient in values approaching 0.5 and enriched in values equal to or approaching 1.0 in order to identify non-recombinant regions. Non-recombinant regions, called "SNP deserts," are multi-megabase areas of the genome with very few sites of discordant chastity ≈0.5, selected for their enrichment of sites matching the reference (ChD = 0), but containing SNPs with discordant chastity approaching 1.0, which likely indicate causative mutations. SNP deserts, then, are the most likely areas to contain the causative SNP of interest, provided sufficient sequence coverage depth, since they show enrichment for one ecotype due to selection on the mutant trait.

A variety of other mapping by sequencing tools are available for use – however, many tools have limitations owing to requirements for large numbers of F2 progeny, prior genetic mapping, or extensive backcrossing [53-55]. A novel technique, which circumvents these limitations (among others), called NIKS (Needle in the K-Stack), is a reference-free algorithm which does not require segregating populations or previously generated genetic maps [56]. NIKS can directly compare sequence reads acquired from mutant and wild-type plants to detect causative mutations. Mutations underlying phenotypes of interest can be directly identified by sequencing mutants and performing a direct comparison with wild-type genomes. A shared characteristic of mapping by

sequencing, whether done through commercially available software packages or custom programs, has been the need for a reference sequence to which re-sequenced reads could be aligned. Despite a multitude of genome-sequencing projects underway to develop reference genome sequences beyond those available for model organisms, reference-based mapping still holds drawbacks. Genomes are not static, and especially fast-evolving genes can change and not be represented in the reference sequence – necessitating re-sequencing of the reference to keep up with genomic evolution. In NIKS, the sequence reads from two highly related genomes are compared to each other, for example a wild-type Col-0 plant versus a phenotype displaying Col-0 mutant. NIKS uses k-mers, which are subsequences of length k of sequencing reads. K-mers specific to each sample, wild-type or mutant, are identified and merged into longer sequences, called seeds. NIKS then creates "seed pairs" by matching up wild-type and mutant seed sequences. Seed pairs which are distinguished only by a mutagen-induced mutation are then separated, and NIKS creates local *de novo* assemblies surrounding the mutated locus using all overlapping read pairs. These extended seeds, or contigs, are several hundred base pairs in length and include the mutated site. Genetic and physiological relevance of putative SNPs can then be determined for each of the contigs – functional significance of genes can be determined by BLAST and homology modeling. An assessment of NIKS in both mice and maize found heightened sensitivity in unique regions (>90%). With more than 25X genomic coverage, >98% of predicted mutations were correctly identified. Drawbacks of the software include its inability to identify mutations within repetitive regions and a high false positive rate.

1.9 Thesis Statement

Permanent root-growth arrest of *atr*$^{-/-}$ Arabidopsis plants in response to UV-B irradiation is dependent on another gene, *ursu*. Previously, EMS mutagenesis was used to derive the *sog1* mutant line. We believe the mutation responsible for the root termination phenotype resulted from the same EMS mutagenesis treatment that created the *sog1-1* mutation. The unidentified mutation is likely a hitchhiker that was unintentionally combined with *atr* in attempts to create the *atr sog1* double mutant.

This hypothesis leads to two predictions. First, if the root-termination phenotype is dependent upon a mutation in an additional gene (*ursu*), this mutation should be identifiable using a next-generation sequencing and bioinformatics mapping approach. Existing bioinformatics tools, such as Next-Generation Mapping (NGM) and Mapping and Alignment with Short Sequences (MASS), have been used in the mapping and identification of causative mutations and should reveal two readily apparent deserts or peaks, respectively, in their histograms – one for *atr* and one for *ursu*. Second, if *ursu* in combination with *atr* is responsible for root termination, once a mutation has been identified, the phenotype should be reversed in complementation tests. To test these predictions, I use high-throughput genome sequencing and bioinformatics to analyze the mutant genome for causative mutations.

Materials and Methods

2.1 Plant Lines

Arabidopsis thaliana seeds (ecotype Col-0) containing an *Agrobacterium tumefaciens* T-DNA loss-of-function insertion in exon 10 of *atr* were provided by Kevin Culligan (University of California, Davis, CA) and were previously characterized [36, 57]. Col-0 seed with a loss-of-function, EMS generated point mutation in the DNA binding domain of SOG1 was provided by Anne Britt (University of California, Davis, CA) and was also previously characterized [40, 41, 58].

The *atr-2 ursu5* line was discovered by Robert Ursu in the Hays laboratory during his undergraduate studies and is unpublished. *Ursu5* was discovered while screening a segregating F2 population derived from a cross between *atr-2* and *sog1-1* according to root growth and stem cell assays described by Curtis et al. [59]. Among UV-irradiated (0.03 kJ m^{-2} UV-B) segregates, several seedlings (18/200) showed a novel irreversible termination of primary root growth (Robert Ursu, unpublished). Previous experiments involving UV-B irradiation of *atr* single mutants and wild-type Arabidopsis revealed transient arrest of root growth, but recovery of root growth was consistently observed within six days [40, 59]. Genotyping PCR verified ATR disruption in all phenotype-displaying mutants, although a terminator with the wild-type sog1 locus was also discovered (unpublished). This plant line, $atr^{-/-}$ $sog1^{+/+}$, was called *atr-2 ursu5* and is the line used for mapping *ursu*.

To initiate mapping experiments, the pollen of an *atr-2 ursu5* plant (ecotype Col-0) was rubbed onto the stigma of a wild-type Landsberg *erecta* (Ler) ecotype mapping line to

generate heterozygous, hybrid F1 seed. Like Col-0, Ler is a well characterized, heavily studied Arabidopsis ecotype. Among the valuable resources developed for genetic analysis of Arabidopsis are extensive collections of marker and polymorphism data. Landsberg *erecta* was chosen as the mapping line because there are a number of naturally occurring single nucleotide polymorphisms (SNPs) between wild-type Ler and Col-0 [60-62]. The single nucleotide differences between the ecotypes have been previously identified. The recessive trait was not identifiable by phenotypic screens in the heterozygous F1 population, and so the F2 seed was collected following self-pollination of the F1 plants to generate a segregating F2 population. Approximately 800 segregating F2 seed were surface sterilized, sown on agar growth medium, and screened for individuals that terminated primary root growth, according to procedures outlined by Curtis et al. [59]. Segregation ratios, obtained from segregation analysis, yield information about the monogenic (or digenic, trigenic, etc.) nature of a trait. The χ^2 statistic, used to determine if segregation ratios match hypothetical Mendelian segregation hypotheses, is calculated using differences in expected and observed numbers of progeny displaying parental and recombinant phenotypes, where expected numbers of progeny are calculated based on Mendelian segregation ratios of two recessive alleles segregating independently in a population in Hardy-Weinberg equilibrium:

$$\sum \frac{(observed - expected)^2}{expected}$$

Significance can be evaluated by converting χ^2 into a p-value. Segregation analysis of F2 individuals revealed a dependence of the growth arrest phenotype on an unidentified gene, referred to as *ursu*, which only displayed the phenotype in the absence of functional

ATR (unpublished). The root termination phenotype appears to occur uniquely in *ursu atr* double mutants, but not in UV-B irradiated *atr* single mutants, and does not depend on the *sog1-1* mutation (unpublished). Sixty-four whole seedlings showing root termination were collected and frozen at −80° C.

2.2 Sequencing Library Preparation

DNA was extracted from whole plants using the Qiagen DNeasy Plant Mini Kit (Qiagen Inc. Germantown, MD) as per manufacturer recommendations. DNA was quantified using a NanoDrop 1000 UV-spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) and PicoGreen quantitation assays (Life Technologies, Carlsbad, CA), and equal masses of DNA from the 64 plants were pooled. Pooled DNA was sheared by sonication to generate molecules between 200 bp and 1000 bp. Sheared DNA was diluted to 3.1 ng/µl using Amicon 100 kD microconcentrators (Millipore, Billerica, MA). A paired-end whole genomic DNA sequencing library was prepared with 154 ng DNA using the TruSeq DNA Preparation kit as per Illumina TruSeq manual revision C. Oligonucleotide adapters containing one of 24 barcodes were used in preparation of the pooled plant DNA to track and identify this sample, which was included with others in five lanes of multiplex, paired-end, 101-bp sequencing using the Illumina HiSeq 2000 (Illumina Inc., San Diego, CA). DNA was pooled from five identical PCR library amplifications and submitted for sequencing to the Center for Genome Research and Biocomputing (CGRB) at Oregon State University. Over two million short sequence reads were generated by the Illumina HiSeq 2000 sequencer.

2.3 Bioinformatics

Sequence reads were aligned to the TAIR10 Col-0 *Arabidopsis thaliana* reference genome using Burrows-Wheeler Aligner (BWA) [63]. The BWA-short algorithm was used to first index the TAIR10 reference in FASTA format, then to align the sequences against the reference, and finally, generate alignments in the sequence alignment/map format for use with SAMtools software version 0.1.13 [64]. Data was converted from sequence alignment/map format to its binary equivalent (BAM) format and indexed again for SAMtools pileup to generate a pileup of read bases from the alignment against the reference. SNP calls from SAMtools were then preprocessed and mapped using Next-Generation EMS mutation mapping (NGM) software [49]. NGM is a validated, efficient method to map mutations using advanced genomic technology with as little as a single lane of flow cell data [49]. During preprocessing, SNP data was filtered using the following default quality filter criteria: minimum depth, maximum depth, minimum neighbor quality, minimum best read quality, and minimum consensus quality. The SNP frequencies in the mapping population, Ler, are binned at 250 kb intervals by default in a histogram displaying the full Arabidopsis genome. As expected, there is a certain degree of natural variation across each chromosome, resulting from Ler-specific polymorphism produced from the Ler x Col-0 cross. The non-recombinant area containing the mutation of interest does not follow this pattern. Instead, this region creates a SNP desert, a multi-Mb area with little to no SNPs. SNP deserts were identified and used to narrow the list of suspect causal mutations. NGM uses a chastity statistic, termed discordant chastity (ChD), to distinguish between SNPs arising from natural variation and those potentially responsible for causative mutations. ChD is a ratio comparing the number of reads that

support either allele of a SNP. A signal corresponding to freely segregating (i.e. no selection), natural variation has a ChD of approximately 0.5 (each allele of polymorphism is equally supported by sequence reads), whereas a signal corresponding to mutation has a ChD of approximately one. Refined estimates of the mutation position are then obtained by ratios of ChD values associated with natural variation and with mutation. Combinations of synonymous substitutions, identical mutations in splice variants, transversion mutations, and non-CDS mutations were removed to generate a final list of candidates in regions identified as SNP deserts.

To complement and verify the results of NGM, the Mapping and Alignment with Short Sequences (MASS) pipeline of algorithms was also used to analyze the short sequence reads from Illumina. A list of 305,002 known homozygous Ler SNPs relative to Col-0 was retrieved from: ftp://ftp.arabidopsis.org/Polymorphisms/Ecker_ler.homozygous_snp.txt (November 2012). A database of SNP clusters – 201 bp sequences centered on each SNP – was created. When SNPs were within 101 bp of each other, these clusters extended 101 bp out from the SNP furthest out to create a larger cluster. A Cache Assisted Hash Search with Xor logic (CASHX) database was created for both the clustered SNPs and the TAIR10 Arabidopsis genome before working through the pipeline [51]. The processing program within the pipeline maps the reads to the genome excluding repetitive regions and low complexity sequences (repeat masked), parses out the quality and sequence information from FASTQ files, and aligns reads across SNP sites using CASHX. Once data was run through MASS Processing programs, analysis algorithms parsed hits to a manually specified mapping interval (~1-2 Mb), mapped reads using Mapping and Assembly with

Quality (MAQ), and filtered SNP lists created by MAQ. Four intervals on chromosome five were selected for further analysis based on the location of the tallest peaks in the histogram: 8.5 - 10.5 Mb, 14.5 - 16.5 Mb, 16.5 - 18.5 Mb, and 18.5 - 20.5 Mb. The list of SNPs for each interval was examined in 100-kb bins and putative causal mutations recorded and further investigated for their physiological implications.

2.4 Validation PCR and Sanger Sequencing

NGM and MASS analysis each revealed a spectrum of potential gene candidates. Gene candidates were pursued and ranked by their physiological relevance. PCR amplification and Sanger sequencing (Center for Genomics and Biocomputing) were performed for validation of one putative causal mutation, *arf13*. Oligonucleotide primers were designed using Primer3 with default parameters (Table 1) [65]. DNA from two *ursu* and two wild-type plants was selected to validate the SNP. A ~200 bp fragment surrounding the mutation of interest was amplified using PCR in a 10 μl reaction containing 1X buffer (500 mM KCl, 100 mM Tris HCl, 15 mM MgCl2, and 1% Triton X-100) (GenScript, Piscataway, NJ), 250 μM each dNTP, 0.5 units Taq polymerase, 1 μM forward and reverse primers and cycled for an initial denaturation step at 98 °C for three minutes, followed by 30 cycles of: 95 °C for 15 seconds, 52 °C for 30 seconds, 72 °C for 20 seconds, and a final elongation cycle at 72 °C for five minutes. Excess dNTPs and primers were removed after PCR and prior to sequencing using Qiagen QIAquick PCR purification kit (Qiagen Inc. Germantown, MD). DNA concentration and quality was confirmed using the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). PCR products were directly Sanger sequenced by the Center for Genomics and Biocomputing core laboratory using an ABI Prism 3730 Genetic Analyzer

Table 1. Primers used for Sanger sequencing.

Primer name	**Primer sequence (5' → 3')**
Arf13 F	CCATCTGTCCAATTGAATCATGC
Arf13 R	GCTTGAATGCACTAGTTTCACAT

and BigDye Terminator v. 3.1 Cycle Sequencing Kit (Life Technologies, Carlsbad, CA). CLUSTAL W was used to align the mutant sequences against the reference sequence available from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org) [66] [67].

Results

A root-growth screen of *atr sog1* double mutants revealed irreversible root growth termination in a subset of plants – an aberrant response to UV-B irradiation. Of 200 root tips that were irradiated with 0.03 kJ m^{-2} UV-B, 18 plants displayed the permanent-root-termination phenotype: six days after UV-irradiation, roots did not re-initiate growth. The phenotype required the absence of functional ATR and was agnostic to the function of SOG1 (unpublished, Ursu, R.). For further investigation of the phenotype, root terminator plants homozygous for *atr* and containing wild-type copies of sog1 were used. Segregation analysis of the root-terminator phenotype suggested a role for an unknown mutation (in addition to $atr^{-/-}$), which was designated *ursu*.

To determine if root termination correlated with stem cell death in the RAM, microscope-imaging assays of 15-20 *ursu atr* roots irradiated with varying levels of UV-B were conducted (unpublished, Curtis, M.). Elevated stem cell death was observed at lower UV-B doses (~0.075 kJ m^{-2}) in the *ursu atr* double mutant, as compared to the *atr* single mutant. At higher doses (>~0.15 kJ m^{-2}), stem cell death in *atr* and *ursu atr* mutants is nearly identical (Figure 3).

To map and identify the gene responsible for the root-termination phenotype (*ursu*) in response to UV-B irradiation, a phenotypic *ursu atr* plant (Col-0 ecotype) was crossed with a wild-type plant (Ler ecotype). From 64 root terminating F2 Arabidopsis progeny, a total of 2,171,475 101-bp sequences were obtained using the Illumina HiSeq 2000. Only 154 ng of DNA was available to use to make the sequencing library, and although the amount of DNA was lower than ideal (~1 μg), the library was prepared and submitted to

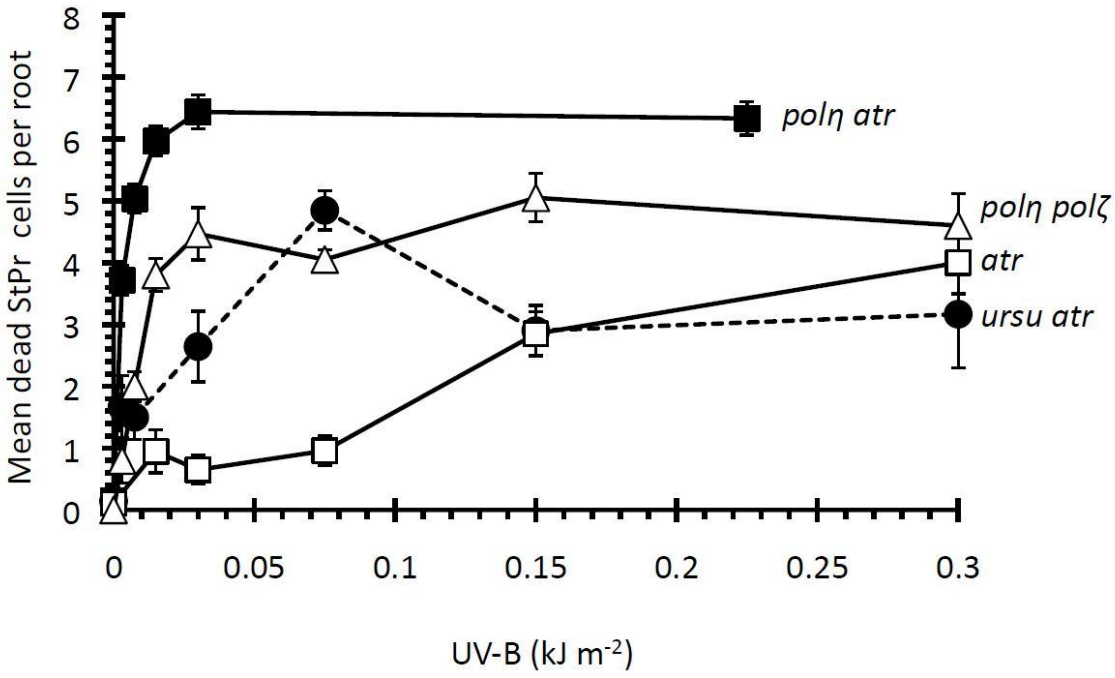


Figure 3. UV-B dose-response curves of *polη*, *polζ*, *atr*, and *ursu atr* mutants. Roots of indicated genotypes were irradiated to indicated UV-B doses, incubated 24 hours, and 15–20 roots each scored for mean stem and progenitor (StPr) cell death (unpublished, Curtis, M.).

the Center for Genomics and Biocomputing for sequencing. Sequence coverage depth of about 1.75X haploid Arabidopsis genomes was obtained (Table 2).

3.1 Next-Generation Mapping (NGM)

The Next-Generation Mapping approach (NGM) generates a histogram plotting density of reads containing non-reference (i.e. non-Col-0) polymorphisms across the genome. The local enrichment of Col-0-specific polymorphisms, identified in the histogram by the absence of signal (i.e. a SNP desert), significantly narrows the search for potential causative mutations. As expected, the *ursu atr* double mutant showed enrichment for Col-0 (i.e. a majority of reads matched the TAIR10 Col-0 reference) around *atr*, creating an obvious SNP desert in the histogram (Figure 4). A second SNP desert that may indicate the presence of *ursu* was not so apparent. Instead, a multitude of "micro-deserts" (~250 kb) were seen throughout the histogram. Genes containing C → T or G → A coding sequence (CDS) mutations in each micro-desert were examined to determine physiological relevance (Table 3). A putative causal SNP was detected in auxin response factor-13 (*arf13*), which fits with a hypothesis for the aberrant growth phenotype (discussed in section 4.6); *arf13* on chromosome one, contained a C/G → T/A SNP at position 12,444,275.

Sanger sequencing and CLUSTAL W analysis of two wild-type and two mutant plants were used to validate the results of NGM (Figure 5). The presence of a C → T mutation was independently verified in each mutant. The expected effects of this mutation varied in each of three possible gene models. For genes where there is conflicting or incomplete information about genetic structure (i.e. exon-intron boundaries), TAIR presents multiple

Table 2. Illumina HiSeq 2000 sequencing results. Data obtained from paired-end sequencing of *ursu atr* indicates low coverage of the Arabidopsis 125 Mb (haploid) genome.

Read length	101 bp
Unique sequence reads	2,171,475
Genome coverage of 125 Mb Arabidopsis haploid genome	1.75X

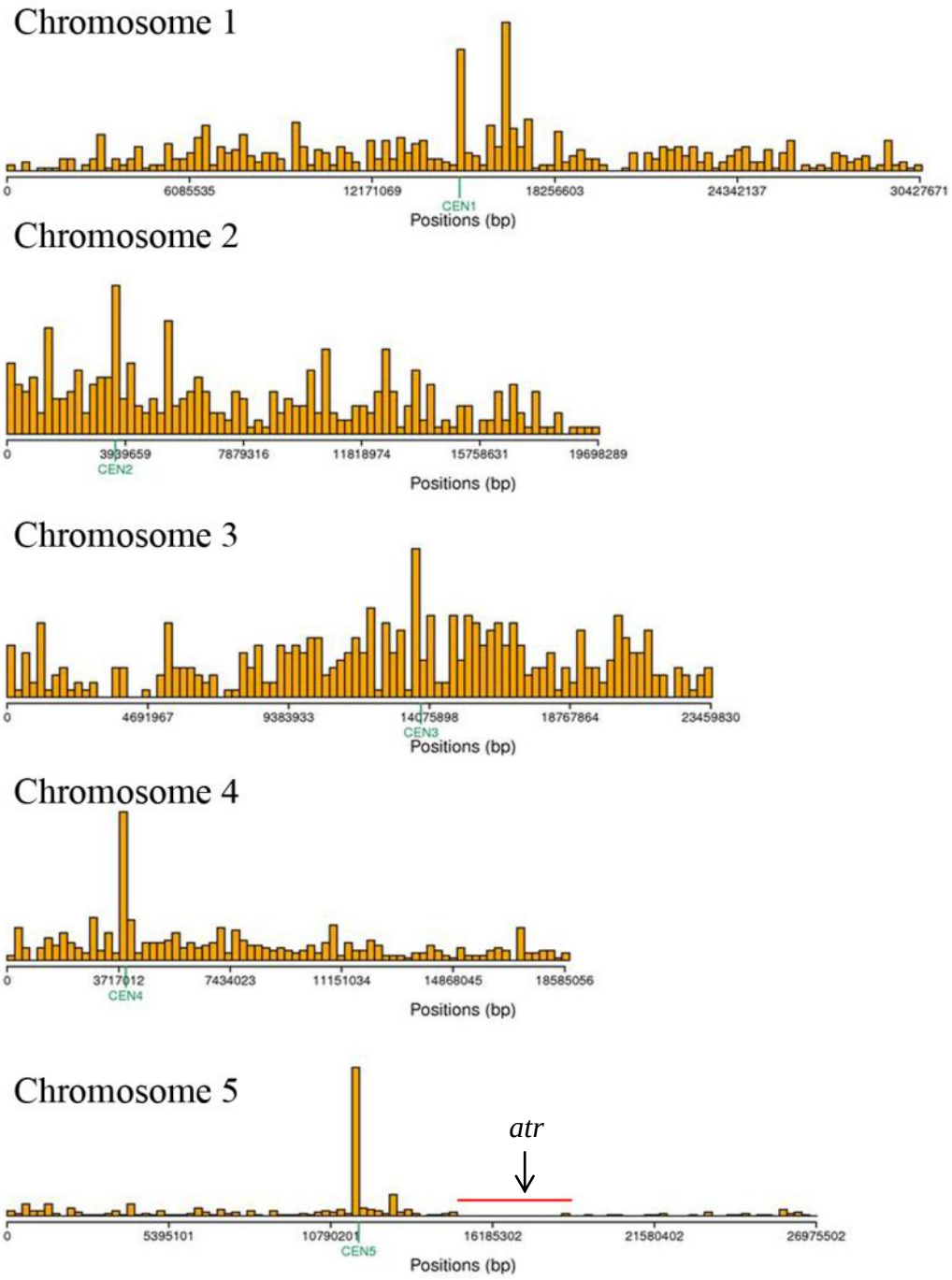


Figure 4. NGM generated genome-wide SNP frequencies for *ursu atr*. SNPs are binned at 250 kb intervals and their abundance is plotted as a function of chromosomal position. SNPs were filtered using the following quality criteria: 6X minimum depth, 100X maximum depth, a minimum best read quality of 20, and a minimum consensus quality of 20. A total of 149,629 SNPs arising from differences between Ler and the Col-0 reference genome were identified. Areas with little polymorphism (SNP deserts) are due to selection on Col-0 traits. The location of the most prominent SNP desert is indicated by a red bar, and the location of *atr* indicated by an arrow.

Table 3. Candidate mutations generated using NGM. Synonymous substitutions, identical mutations in splice variants, transversion mutations, and non-CDS mutations were removed to generate the list of candidates for chromosomes 1-4 and for the SNP at position 7653884 on chromosome 5. The remaining chromosome 5 candidates, in an effort to identify mutations specifically under the multi-megabase SNP desert, were generated by removing only identical mutations in splice variants and non-CDS mutations. The SNP identified in *arf13* (highlighted) was validated by Sanger sequencing.

Chromosome	Position	Reference base	Consensus base	Discordant chastity	Read depth	Feature
1	7798124	G	A	1.00	8	AT1G22100.1
1	9034883	G	A	1.00	14	AT1G26130.1
1	12203196	G	A	1.00	11	AT1G33670.1
1	12444275	G	A	1.00	9	AT1G34170.1
2	5639715	G	A	1.00	8	AT2G13540.1
3	19576347	G	A	1.00	8	AT3G52820.1
4	2960323	G	A	1.00	9	AT4G05612.1
4	7203682	G	A	1.00	8	AT4G12020.1
4	8296650	G	A	1.00	8	AT4G14400.1
4	17094000	C	T	1.00	10	AT4G36120.1
4	17094017	G	A	1.00	10	AT4G36120.1
5	7653884	C	T	1.00	8	AT5G22890.1
5	13604517	C	T	1.00	9	AT5G35405.1
5	26121846	A	C	1.00	10	AT5G65370.1

```
CLUSTAL W (1.83) multiple sequence alignment

ARF-13 WT sequence (TAIR)  -CCATCTGTCCAATTGAATCATGCAAAAAATCAGAAATTTCAAAACTCAA
WT arf-13, sample 1        TCCATCTGTCCAATTGAATCATGCAAAAAATCAGAAATTTCAAAACTCAA
WT arf-13, sample 2        TCCATCTGTCCAATTGAATCATGCAAAAAATCAGAAATTTCAAAACTCAA
Ursu atr, sample 1         TCCATCTGTNCAATTGAATCATGCAAAAAATCAGAAATTTCAAAACTC--
Ursu atr, sample 2         TCCATCTGTCCAATTGAATCATGCAAAAAATCAGAAATTTCAAAACTC--
                            ******** ************************************

ARF-13 WT sequence (TAIR)  AAATCAAAAAGCAACCACTAGTTGCCTCAAGATAAAAAGTTTGACCAAAC
WT arf-13, sample 1        AAATCAAAAAGCAACCACTAGTTGCCTCAAGATAAAAAGTTTGACCAAAC
WT arf-13, sample 2        AAATCAAAAAGCAACCACTAGTTGCCTCAAGATAAAAAGTTTGACCAAAC
Ursu atr, sample 1         ------------------------TCTCAAGATAAAAAGTTTGACCAAAC
Ursu atr, sample 2         ------------------------TCCCAAGATAAAAAGTTTGACCAAAC
                                                   * **********************

ARF-13 WT sequence (TAIR)  CCAACCTCTGAGATCACCAAAAGAGGTCCAAAGCACGGAATTCAATTTTA
WT arf-13, sample 1        CCAACCTCTGAGATCACCAAAAGAGGTCCAAAGCACGGAATTCAATTTTA
WT arf-13, sample 2        CCAACCTCTGAGATCACCAAAAGAGGTCCAAAGCACGGAATTCAATTTTA
Ursu atr, sample 1         TCAACCTCTGAGATCACCGAAAGAGGTCCAAAGCACGGAATTCAATTTTA
Ursu atr, sample 2         TCAACCTCTGAGATCACCGAAAGAGGTCCAAAACACGGAATNCAATTTTA
                            ***************** ************* ******** ********

ARF-13 WT sequence (TAIR)  CTAGAAGTCGTATTAAAGTAAGC-ATAAAATC-ATTATATCTGTAACATA
WT arf-13, sample 1        CTAGAAGTCGTATTAAAGTAAGC-ATAAAATC-ATTATATCTGNNAACNA
WT arf-13, sample 2        CTAGAAGTCGTATTAAAGTAAGC-ATAAAATCTATTAAAATNNTGNANCN
Ursu atr, sample 1         CTAGAAGTCGTATTAAGGTAAGC-ATAAAATC-ATTANATCTGNNACATN
Ursu atr, sample 2         CTAGAAGTCGTATTAAGGTAAGCCATNAAATC-TTTATATCTGTANCNNA
                           *************** ****** ** *****  *** *

ARF-13 WT sequence (TAIR)  TGACTTTTTTTTTAAAATGTGAAACTAGTGCATTCAAGC
WT arf-13, sample 1        TANNAACTTTNNT-------------------------
WT arf-13, sample 2        ANTATAANTNTNNN-------------------------
Ursu atr, sample 1         NNANNNNNNNNN-------------------------
Ursu atr, sample 2         NNTNNANCTNNNNNN-----------------------
```

Figure 5. CLUSTAL W multiple sequence alignment. Wild-type arf13 sequence (obtained from TAIR), corresponding Sanger sequenced wild-type samples, and corresponding Sanger sequenced *ursu atr* mutant samples were aligned. The asterisk (*) symbol below indicates a match among all five sequences, and absence of the asterisk indicates a mismatch in at least one sample. The putative causal SNP identified by NGM and the deletion are outlined in red.

gene models of alternative splicing variants [67]. In gene models one and two, the SNP is in exon 12, whereas in gene model three, the SNP is in intron 12. The SNP was also much closer to the stop site in gene models one and two. In these two gene models, the SNP caused a non-conservative missense mutation: the proline three codons away from the STOP codon was changed to leucine.

Unexpectedly, Sanger sequencing also revealed the presence of a 26-bp deletion, in addition to a C → T SNP (Figure 5). *In silico* translation of gene model one that contains the deletion (in isolation or in conjunction with the C → T SNP) results in a frameshift mutation. This frameshift causes the loss of the original STOP codon and extends the protein length from 505 amino acids (wild-type length) to 625 amino acids. The deletion in gene model two (in isolation or in conjunction with the C → T SNP) is predicted to cause a frameshift mutation which extends the protein length from 479 amino acids (wild-type length) to 599 amino acids. The protein encoded by wild-type gene model three is 546 amino acids in length. In this model, the SNP and deletion fall in an intron and they do not independently or in conjunction disrupt the conserved intron 5' donor site, 3' acceptor site, or the branch (lariat) site.

3.2 Mapping and Alignment with Short Sequences (MASS)

MASS generates plots of the ratio of Col-0 vs. Ler aligned reads along the genome, with a positive read count indicating Col-0 enrichment. Examination of all five Arabidopsis nuclear chromosomes revealed the tallest, most unmistakable peak to be located directly above *atr* on chromosome five (Figure 6). Chromosomes one and three showed moderate enrichment for Ler in certain regions. A plot of ecotype-specific enrichment for chromosome two was relatively unremarkable. Chromosome four contained a

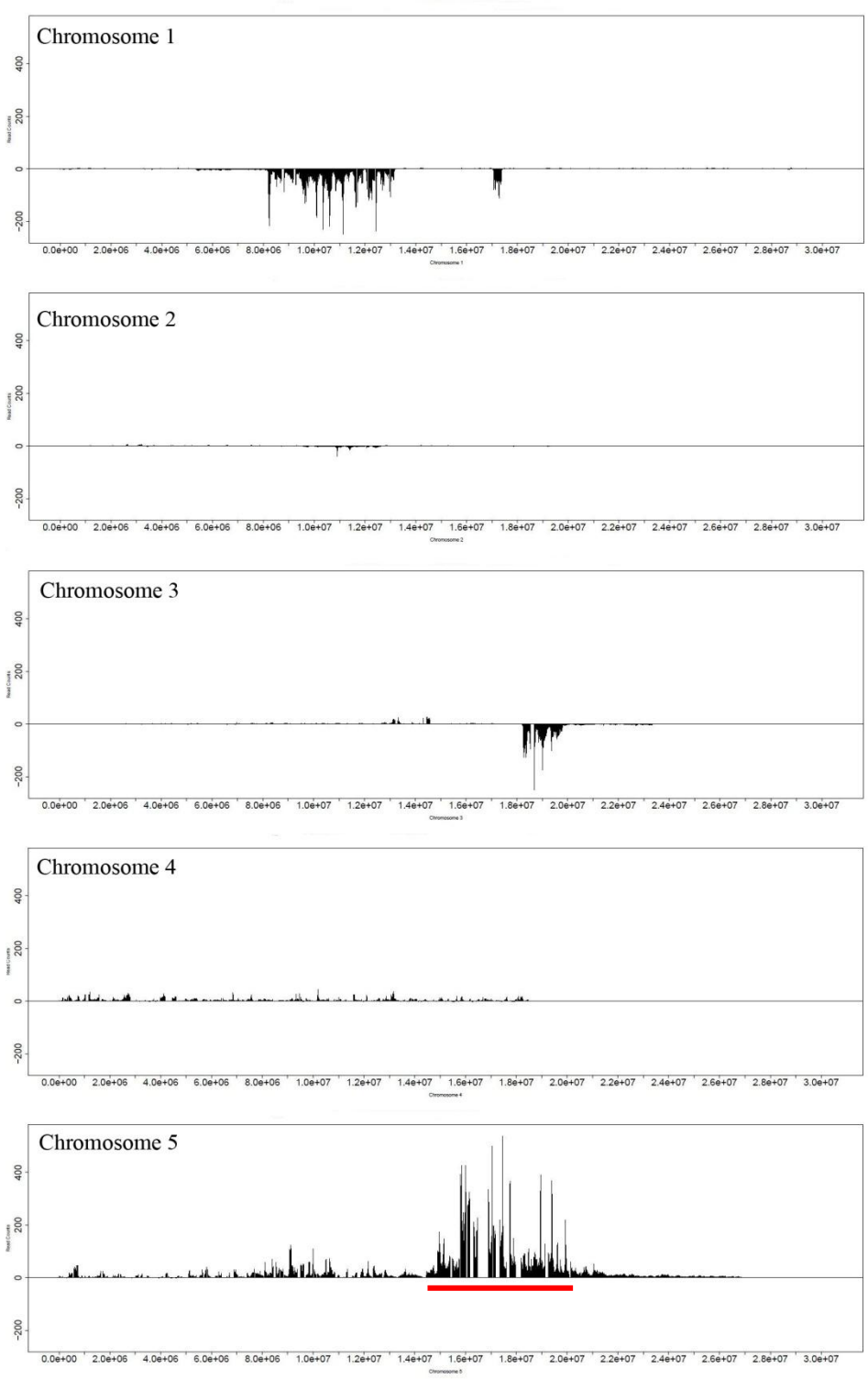


Figure 6. MASS generated genome-wide Col-0 vs. Ler SNP ratios. Read counts vs. base pairs are plotted for *ursu atr* binned at 100 Kb intervals. Positive read counts indicate enrichment for Col-0. The red bar indicates the region likely harboring the causative mutation(s). This indicated region also contains the atr locus.

homogeneous population of short, unremarkable peaks selected for Col-0. The largest signal for Col-0 enrichment was observed on chromosome five, centered on *atr*.

Using MASS involves manually selecting a mapping interval (~1-2 Mb) centered on peaks of interest. Four 2-Mb intervals on chromosome five were selected for further analysis using MASS: 8.5 - 10.5 Mb, 14.5 - 16.5 Mb, 16.5 - 18.5 Mb, and 18.5 - 20.5 Mb. Hundreds of SNPs were discovered within each mapping interval, but only those immediately under or adjacent to areas underneath large peaks were considered. Of the SNPs that fit these criteria, only SNPs in positions that disrupted non-transposon, protein coding and RNA genes were considered to be putative, causal mutations. The refined list of potential SNPs fell in loci containing small RNAs (smRNAs) and protein coding genes (Table 4).

Under the peak representing the largest enrichment of Col-0 DNA (largest ratio of Col-0-specific sequence reads to Ler-specific sequence reads), between 16.5 - 18.5 Mb along chromosome 5, only one mutation was detected in the 100-kb bin represented by the tallest histogram peak. The T/A → C/G SNP (position 17466479, chromosome five) is located in the first exon of the gene for Resistance to *Peronospora parasitica* protein 8 (rpp8). *In silico* translation revealed the SNP is a synonymous mutation in the wobble position of the 60th codon (Asparagine 60).

Table 4. Candidate mutations generated using MASS. Shown are relevant mutations in the *atr*-linked region and on chromosome five.

Chromosome	Position	Reference base	Consensus base	Phred-like consensus quality	Read depth	Feature
5	9093913	A	C	42	5	smRNA
5	15844893	T	G	36	3	AT5G39570
5	15964369	C	T	36	3	smRNA
5	17085645	T	C	36	3	smRNA
5	17466479	A	G	36	3	AT5G43470
5	19418994	C	G	36	3	smRNA

Discussion

Root growth arrest in *Arabidopsis thaliana* following UV-B irradiation and photoproduct accumulation putatively results from non-functional, mutant *atr* and another unknown gene, *ursu*. Here, I undertook to map the genomic location of the causative mutation, *ursu*. Several putative candidates were identified, including *arf13* and *rpp8*, but ultimately, the search for a causal mutation was inconclusive.

4.1 Rationale

The initial, broader goal of mutagenizing and combining DDR mutants, including *sog1* and *atr*, was to elucidate the contribution of different DDR proteins in response to a model mutagen, UV-B. Prior to the discovery of the root termination phenotype, the study followed a traditional reverse genetics approach: knocking out essential proteins involved in the DDR, creating double or triple mutants through crosses, and observing responses following UV-B irradiation. A *sog1* mutant was created by EMS mutagenesis – a popular technique known to generate a random distribution of mutations throughout the genome [68]. An EMS-derived mutant *sog1* was crossed with an *atr* (T-DNA insertion knockout) deficient plant to create the *atr sog1* double mutant. A unique phenotype emerged following UV-B irradiation of the segregating F2 *atr sog1* roots: permanent root growth arrest. PCR genotyping of these plants revealed dependence of the phenotype on *atr*, but not on *sog1* deficiency. Furukawa et al. [40] had previously performed similar UV-B root growth assays using *atr* plants, and had observed no such phenotype. Segregation analysis suggested the dependence of the phenotype on a recessive gene, in addition to the lack of *atr*. At this point, the experiment switched to a forward genetics approach in order to identify and map the gene responsible for this unique phenotype.

4.2 High-Throughput Sequencing

To identify the location and identity of the mutant gene causing the *ursu* phenotype, we undertook to map the gene using a mapping by sequencing approach. After crossing the Col-0 *ursu* mutant with the Ler mapping-line we sequenced the DNA from plants that displayed irreversible root growth arrest after UV-B irradiation.

DNA from 64 F2 mutant plants was extracted, pooled, sheared, and used to create a sequencing library for mapping by sequencing. A total of 2,171,475 unique 101-bp sequence reads were produced by the Illumina HiSeq 2000 for a dismal 1.75X coverage of the 125-Mb Arabidopsis haploid genome. Overlapping reads generated by the sequencer create coverage across the genome, where coverage is defined as the average number of reads covering any nucleotide [69]. Multiple reads per base are necessary for reliable base calling. Also, because reads are not distributed evenly across the genome, some regions will experience more or less coverage than the average. Deeper sequence coverage is necessary to ensure that all regions of the genome are covered by some overlap of reads. Next generation sequencing projects, whether or not mapping by sequencing, commonly aim for 25X – 50X haploid genomic coverage [54, 70, 71]. A recent study of the effects of genome coverage on mapping by sequencing experiments recommends a minimum of 25X coverage in order to avoid overlooking causative mutations [72].

DNA from 64 F2 progeny of the cross between *ursu atr* (Col-0) and wild-type Ler with permanently arrested root growth following UV-B irradiation was sequenced. Although there were enough individuals to use MASS or NGM to determine loci where Col-0 DNA was enriched by selection (MASS and NGM have been used successfully with 92 and 10

individuals, respectively), low sequence coverage of the genome was acquired. DNA extraction from cotyledons of the 64 individuals yielded little DNA in large volumes. Each sample was eluted in large volumes to maximize yield (100 μl), and the extractions yielded enough DNA, once pooled, for sequencing. An empirical determination of shearing time was performed using a different sample, which was to be sequenced in parallel with the *ursu* sample. Following sonication and concentration of each sample, 154 ng *ursu* DNA was available for sequencing library preparation. This loss of DNA during the shearing step could be due to adhesion of DNA molecules to the walls of microfuge tubes. The more probable reason is, due to a difference in concentration which was not taken into account, the DNA was over-sheared during sonication (due to being sonicated for too long and in larger volumes). Compounded with potential overshearing of DNA, Amicon microconcentrators removed fragments <100 bp and so more DNA could have been lost in this manner. To avoid these issues, the pooled DNA should have been concentrated with the Amicon microconcentrator before shearing. If DNA concentrations were still significantly different than those of the other samples, an empirical test to determine sonication time should have been run independently for this sample. Regardless, several features stood out in this slim dataset.

4.3 Mapping and Alignment with Short Sequences

Mapping analysis of the first four chromosomes of Arabidopsis using MASS was unremarkable, aside from the noted large Ler-specific negative enrichment peak on chromosome one. No regions showed significant enrichment for Col-0 DNA across chromosomes one through four. Chromosome five displayed a salient, Col-0-specific DNA enrichment peak in the region of *atr*. Further analysis was performed in 2-Mb

increments across the length of the several-megabase peak. Candidate SNPs in each 2-Mb interval were filtered manually by functional annotation to assess the potential of each mutation to cause the *ursu* phenotype. A list of probable SNPs in non-transposon, protein or RNA encoding genes was developed. A filtered list of potentially causative SNPs contained only SNPs in four small RNA (smRNA) and two protein encoding genes.

SNPs in smRNAs have been mapped and shown to have phenotypic ramifications [48]. Small RNAs are short RNA sequences (~21-26 nt) responsible largely for repressing gene expression in most eukaryotes through binding to complementary sequences [73]. Small RNAs come in many varieties, including short interfering RNA (siRNA), small temporal RNA (stRNA), heterochromatic siRNA, tiny non-coding RNA, and microRNA (miRNA). In addition to mediating eukaryotic genetic regulation, smRNAs can silence mRNA and target it for degradation, or inhibit its translation altogether. Small RNAs play a role in host defenses to viruses and transposons, and they can target epigenetic modifications to certain areas of the genome. If the causal mutation for the root termination phenotype is indeed in a small RNA, a complementation test transforming *ursu atr* with the wild-type allele of the smRNA in question should restore root growth following UV-B irradiation. Of the smRNAs containing SNPs discovered by MASS, further analysis revealed a multitude of overlapping smRNAs. Without additional sequencing to increase coverage depth, or complementation analysis, it is difficult to further narrow the list of candidates.

Of the two SNPs found by MASS in protein coding genes, one is in the first exon of a gene of unknown function. There is some microarray evidence of root expression, but it is difficult to speculate on the putative effects or pathways by which it would cause the

mutant phenotype [74]. The second SNP found in a protein coding gene was found in the first exon of Resistance to *Peronospora parasitica* protein 8 (rpp8). Although immediate physiological relevance was difficult to muster, the validity of this SNP was strengthened by its presence directly underneath the tallest Col-0 enriched peak. Rpp8 is one of many conserved resistance (R) genes present in plants activated by pathogens [75]. Resistance genes are involved in the recognition of viral, bacterial, fungal, and oomycete pathogens. Rpp8 confers resistance to oomycete pathogen *Peronospora parasitica.* Once the presence of the pathogen has been recognized by resistance genes, defense responses such as the accumulation of salicylic acid, expression of pathogenesis-related genes, and hypersensitive response become activated. In the hypersensitive response, necrotic lesions are formed at the site of infection and a barrier of dead cells is thought to help prevent further pathogen multiplication and movement by limiting nutrients and other necessary host components.

In silico translation of rpp8 containing the SNP reported by MASS resulted in a synonymous mutation (N60 → N60), due to the presence of the SNP in the wobble position of the codon. Sanger sequencing of rpp8 should be performed to verify that there are not other SNPs or insertion/deletion mutations in the gene not detected by MASS. Although the SNP detected is not predicted to have significant effects, it may have been reported by MASS because it may be tightly linked to a deleterious, causative mutation lacking sufficient sequence coverage to identify. Alternatively, this so-called "silent mutation" could fall in the category of an increasing collection of genes in which synonymous mutations appear to have not so silent effects [76, 77]. Synonymous codons, although coding for the same amino acid, can be differentially preferred and used.

Synonymous codon usage bias exists in Arabidopsis, among other model organisms [78]. Increasing amounts of evidence suggests that synonymous mutations experience selective pressures linking them to transcription, splicing, DNA secondary structure, and mRNA secondary structure and stability [77]. Strong selective pressures on synonymous codons exist for translational efficiency and, as a result, highly expressed genes are encoded primarily by codons corresponding to highly abundant tRNAs. By the same token, the presence of an uncommon codon (with less available tRNAs) can slow protein manufacture and affect efficient folding potentially altering rate of translation [79]. Kimchi-Sarfaty et al. [76] saw differences in mRNA and protein conformations between wild-type and synonymous mutant genes. They hypothesized that a synonymous change from a more frequent codon to a rare codon will affect the timing of co-translational folding. As more protein product continues to be made and tRNA becomes depleted, the usage of a particular codon suddenly becomes of importance. In rpp8, codon usage at the identified synonymous SNP (AAT → AAC) decreases only slightly from 23.1% to 20.7%. The differences in frequencies reported by Kimchi-Sarfaty et al. [76] as a result of the synonymous SNP were at least 8%. The 2.4% difference in codon usage is not significant enough in the case of N60 → N60 to warrant further study of this basis. Sanger sequencing is needed to verify the presence of the SNP, and has potential to reveal other nearby causal SNPs or insertion/deletion mutations, which may not have been detected by software due to low coverage and software limitations.

A chromosome plot of the SNPs identified by Joseph Ecker's group and used for MASS was generated (Figure 7). A visual representation of the SNPs across the five chromosomes illustrates a low SNP density reported across the length of chromosome

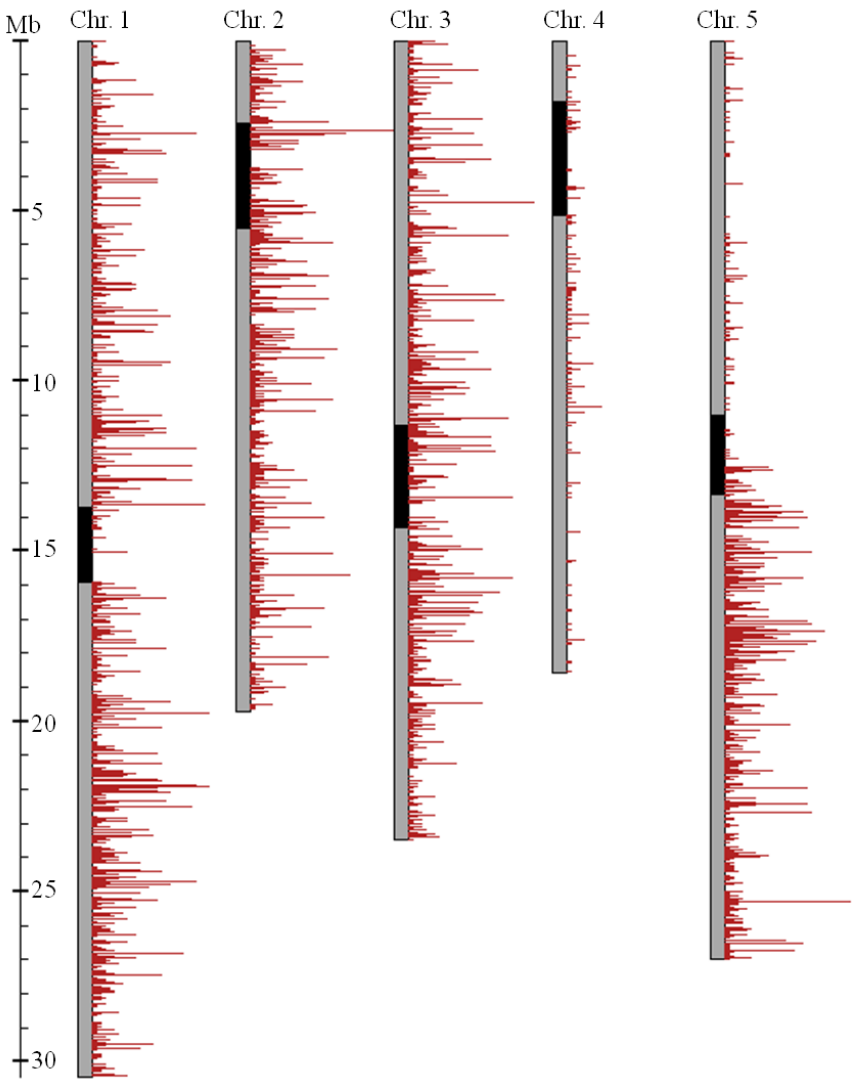


Figure 7. Single nucleotide polymorphic sites between Ler and Col-0 genomes. SNPs are binned every 1000 bp. Chromosome four and the short arm of chromosome five show a lower density of SNPs compared to other genomic regions.

four and along the short arm of chromosome five. Due to overall low sequencing coverage, and lack of two salient peaks among the five chromosomes to explain the phenotype, there emerged three possibilities regarding the location of the causative mutation: (1) the sequencing coverage is far below the threshold required for accurate mutation detection by MASS. This problem is compounded by the low density of SNPs on chromosome four and on the short arm of chromosome five; this SNP list is central to the function of MASS and a low density of SNPs combined with extremely low coverage in these areas holds high potential for causal SNPs to be overlooked. This possibility seems unlikely due to the strong signal in the region of *atr*, an internal control. (2) The causative mutation is tightly linked to *atr*, and the causal mutation is located under the large peak on chromosome five. (3) The phenotype is caused by quantitative trait loci rather than by a qualitative trait locus, as has been assumed. Additional sequencing coverage across this peak would greatly facilitate mapping and identification efforts.

4.4 Next-Generation Mapping

NGM generated a singular, large SNP desert on chromosome five. This desert was initially dismissed from the pool of potential causal SNPs – it was regarded as the signal due to the internal positive control, *atr*. As such, the search for causal mutations began along the other four chromosomes with the assumption that poor coverage was limiting the program's ability to draw another true, multi-megabase SNP desert. It was through the examination of these here called micro-deserts that a SNP in *arf13* (locus AT1G34170) was discovered. Further examination of SNPs located in the SNP desert on chromosome five, even with permissive filtering criteria, did not return salient candidates (Table 3). Physiological relevance of each of the putative causal SNPs was examined,

and of all candidates, *arf13* appeared to have the most compelling case. For these reasons, despite little supporting evidence from the NGM histogram, the *arf13* SNP was investigated further. Sanger sequencing confirmed the presence of the SNP (P → L) detected by NGM in *arf13*, and also revealed a 26-bp deletion. In gene models AT1G34170.1 and AT1G34170.2, the SNP and deletion are both in exon 12. The frameshift-causing deletion, in both gene models, resulted in the loss of the original STOP codon and extension of the protein length by 120 amino acids. In gene model AT1G34170.3, the SNP and deletion are located in intron 12 and do not disrupt conserved intron elements (Figure 8).

Each gene model is ranked according to confidence scores, which are generated based on transcript data, proteomics data, alignments with other plant species, and multiple alignment analysis. For each gene model, individual exon confidence ratings are first generated based on available data of the aforementioned data types. Gene model confidence rankings are constructed from the combined confidence scores of all its exons, as well as from overall rankings based on all available experimental datasets in combination. A ranking out of five stars is given to each gene model, with five stars being the best ranking and representing coverage across the length of the gene and at every splice junction by a single piece of evidence. Gene models AT1G34170.1 and AT1G34170.2 both meet the requirements for five star rankings. Gene model AT1G34170.3 receives only two stars – meaning although there is enough experimental and computational evidence covering >50% of the gene, the evidence does not span every splice junction. Support for gene models AT1G34170.1 and AT1G34170.2 lends credence to a potential effect of a

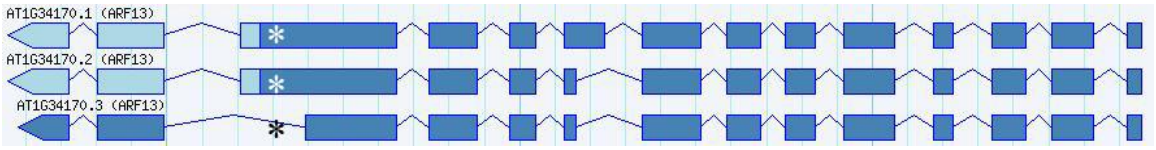


Figure 8. ARF13 gene models presented by TAIR. From top to bottom: AT1G34170.1, AT1G34170.2, and AT1G34170.3. The putative causal SNP is located in exon 12 in gene models AT1G34170.1 and AT1G34170.2 (shown by a white asterisk), and in intron 12 of gene model AT1G34170.3 (shown by a black asterisk).

frameshift mutation in the 12th exon of *arf13*, although a complementation test needs to be performed to verify the causal nature of these mutations on the root termination phenotype. A number of limitations of the gene model ranking approach exist: abundance of supporting data is not taken into account, certain data types automatically generate lower rankings for specific splice sites, models are not penalized for missing elements, all evidence types are taken into account equally, proteomics data uses the map position of the peptide but not the frame, and the ranking system does not provide a sense of the accuracy of the annotated open reading frame [67].

4.5 Auxin Response Factor 13

A large deletion found in *arf13* (auxin response factor 13) causes a frameshift mutation that extends the protein. This alteration may affect the normal functioning of this protein. Auxin response factor 13 is an attractive candidate because of its involvement with auxin signaling and regulation of polarized root growth.

Auxin, an essential phytohormone, plays crucial roles in regulating cell division, extension, and differentiation. Together with other phytohormones, auxin is believed to hold a critical role in nearly every aspect of plant growth and development [80]. Once synthesized, auxin is transported to certain areas of the plant to trigger signaling cascades resulting in developmental responses. Auxin regulates transcription of a variety of genes, including the Aux/IAA, SAUR, and GH3 gene families, many of which include early auxin response genes that are activated rapidly after auxin treatment [81]. Auxin response genes are regulated by two transcription factor families: auxin response factors (ARFs) and Aux/IAA repressors. Auxin response factors (ARFs) are sequence-specific DNA-binding transcription factors that bind to auxin response elements (AuxREs) in the

promoter regions of early auxin response genes such as Aux/IAA [82]. The leading model posits that once auxin is detected by members of the Transport Inhibitor Response 1 family, Aux/IAA repressor proteins are proteolyzed, thereby releasing their inhibitory effects on ARFs. ARFs either repress or activate auxin response genes depending on cell type, environment, and the presence of an activation or repression domain in the variable central region of the protein [83].

Developmental processes in the cell are initiated and mediated through auxin-regulated gene expression [84]. Root development is heavily influenced by precisely organized gradients of auxin [85]. Variation in auxin concentration and auxin responsiveness dictates auxin-regulated activities throughout the plant, including the root. In the growing root tip, auxin concentrations increase along the root toward the root apical meristem, reaching a sustained maximum at the quiescent center (QC) [86]. This particular pattern of auxin in the root tip defines the identity and position of cells in the stem cell niche. The auxin gradient guides cells to transition from the zone of cell division in the root tip (meristem) to the zone of cell elongation higher up the root tip. Auxin is transported directionally toward the root apex by the PIN (PIN FORMED) family of polar transport proteins [87]. At the root apex, auxin is redirected to the lateral root cap, epidermis, and back up toward the elongation zone. Some auxin is recycled back toward the root apex by a subset of PIN proteins, thus creating a circulatory route of auxin transport in the root tip [88]. PLT (PLETHORA) transcription factors are crucial in defining the QC and the stem cell niche in the root apical meristem [89]. In order to generate a specific auxin distribution, PLT genes control several members of the PIN family of proteins. PIN genes are needed to establish the auxin maximum at the QC and since PLT expression is auxin

inducible, PLT expression is thereby localized, specifically, to areas of high auxin concentration. There is evidence of ARF regulation of PLT gene expression [89]. Transcription of auxin transport proteins, including PINs, was also discovered to be under the control of ARFs [90]. Regulation of transport protein expression by ARFs fits in with a proposed transcription regulatory loop, involving proteolysis of Aux/IAA proteins under high auxin concentrations and ARF-mediated up-regulation of auxin transport out of the cell [48]. The subsequent reduction in auxin is sensed and results in a reduction of ARF-mediated transcription of transport components. Thus, auxin gradient production is self-reinforcing.

A total of 22 full-length ARFs have been identified in Arabidopsis [91]. Of these, *arf13* was identified using NGM and functional annotation as the most probable candidate mutation behind the root termination phenotype. In previous studies, arf13 exhibited repressor action in transfected plant protoplasts, however, because activation/repression tests for ARFs have relied on transient expression assays in vitro, the possibility remains that the activity of arf13 could change in response to cellular environment or cell type [81]. Little is known about arf13 function in particular; T-DNA disruption of arf13 did not show obvious growth phenotypes [90]. Arf13, along with a horde of adjacent arf genes on chromosome one, is expressed during plant embryogenesis. Expression profiling of arf genes in Arabidopsis (Col-0 ecotype) root tips revealed some expression in the lateral root cap and epidermis [92]. It has been suggested that particular Aux/IAA-ARF pairs determine auxin-dependent developmental processes [93].

4.6 Mechanistic Model for *arf13* Involvement in Irreversible Root Termination

Furukawa et al. [40] previously characterized *atr*$^{-/-}$ root growth and stem cell death in response to UV-B irradiation (0.3 kJ m-2). ATR recognizes RPA-coated ssDNA adjacent to dsDNA. This structure commonly occurs when replication stalls, at telomeres, during NER, or during replication when DNA polymerase is forced to reinitiate downstream of a disruptive lesion [19]. ATR also plays important roles in cell cycle checkpoints and is responsible for the stabilization of stalled replication forks.

Replication fork stalling during S phase can result from endogenous or exogenous lesions in the template DNA – such as from genotoxic chemicals, replication errors, or UV irradiation. Once the replication fork is stalled, it is crucial to stabilize the fork to provide the cell time to repair the lesion or bypass it, so that replication may eventually resume. ATR is among several proteins recruited for fork stabilization. Signals of replication fork stalling are transmitted directly by ATR to cell cycle effector kinase Chk1 to signal cell cycle arrest. Replication and cell division suffer critical delays when lesions are not removed. In the absence of essential proteins such as ATR, the replication fork may not be stabilized, resulting in replication fork collapse, DSB accumulation, and eventual cell death. ATR is also posited to hold more specific roles in response to UV-induced DNA damage. In the presence of UV damage, ATR phosphorylates the NER factor Xeroderma pigmentosum group A (XPA) protein [94]. ATR-dependent phosphorylation of XPA may promote NER repair of persistent DNA damage due to UV irradiation, especially CPDs. ATR is believed to play a role in the regulation of global genome NER, which removes DNA damage from anywhere within the nuclear genome during S-phase [95]. In non-

dividing cells, UV damage signaling – whether or not in the presence of functional NER – is believed to proceed via a common pathway involving ATR [96].

Furukawa et al. [40] observed growth recovery in *atr* roots after UV-B irradiation. Elevated PCD was observed in UV-B irradiated *atr* mutants relative to wild-type controls, with a mean of approximately four dead stem cells per root in response to 0.3 kJ m^{-2} UV-B [40]. Cell death in *atr* mutants was found mostly in StPr cells, apical to the QC. Unpublished results also show elevated stem cell death in *ursu atr* mutants (Figure 3). Elevated PCD in these mutants makes sense, given the crucial roles of ATR in UV damage signaling, fork stabilization, cell cycle signaling, and communication with other DDR pathways for repair initiation. One potential explanation for the root termination phenotype, if the causal mutation is indeed in *arf13*, is that the StPr cells assist in the maintenance of the auxin maximum, and their death due to a lack of ATR when irradiated results in the collapse of the auxin gradient. The auxin gradient is unable to reform because of the absence of ARF13.

Delayed root growth recovery in response to UV-B irradiation (0.3 kJ m^{-2}) in *pol eta pol zeta* mutants appeared to be caused by a transient loss of organization in the root apical meristem, as analyzed by confocal cross-sectional microscopic imaging [59]. A similar loss of meristematic organization, which cannot recover, may be occurring in the *ursu* mutant. ARF13 shows moderate expression in the lateral root cap and epidermis under non-irradiated growth conditions and appears to be most heavily expressed during embryogenesis. Although the exact function of ARF13 during embryogenesis remains unknown, ARF13 could potentially be important to the initial establishment of the auxin gradient in the root apex. Perhaps expression of ARF13 is also necessary during stressful

conditions that involve severe disruptions to the root apical meristem in order to reestablish the auxin gradient loop with a maximum at the QC. ARF proteins are responsible for regulating transcription of auxin transport components (such as PINs) and essential transcription factors (PLTs) that maintain the auxin maximum at the QC and define the stem cell niche. Death of cells responsible for circulating auxin around the root tip disrupts the auxin gradient, and perhaps without ARF13, the gradient and auxin maximum cannot be reestablished. Support for increased stem cell death in the *ursu atr* mutant is unpublished (Figure 3). QC ablation studies have revealed the importance of the auxin maximum in maintaining stem cell identity. Ablation of the QC results in local stem cell differentiation [97]. Under normal conditions, a new distal auxin maximum is established in adjacent cells, thus re-specifying the QC [98]. However, perhaps without ARF13 to coordinate the reestablishment of the auxin gradient, cells in the root tip differentiate prematurely. The gradient of auxin is responsible for maintenance of cell identity and for promotion of growth, both by proliferation in the meristematic zone and by elongation in the elongation zone. Potentially, disruption of the auxin gradient in the absence of ARF13 may prevent the successful re-establishment of auxin gradients – resulting in permanent root-growth arrest in response to UV-B in *atr* mutants.

4.7 Comparison of NGM and MASS

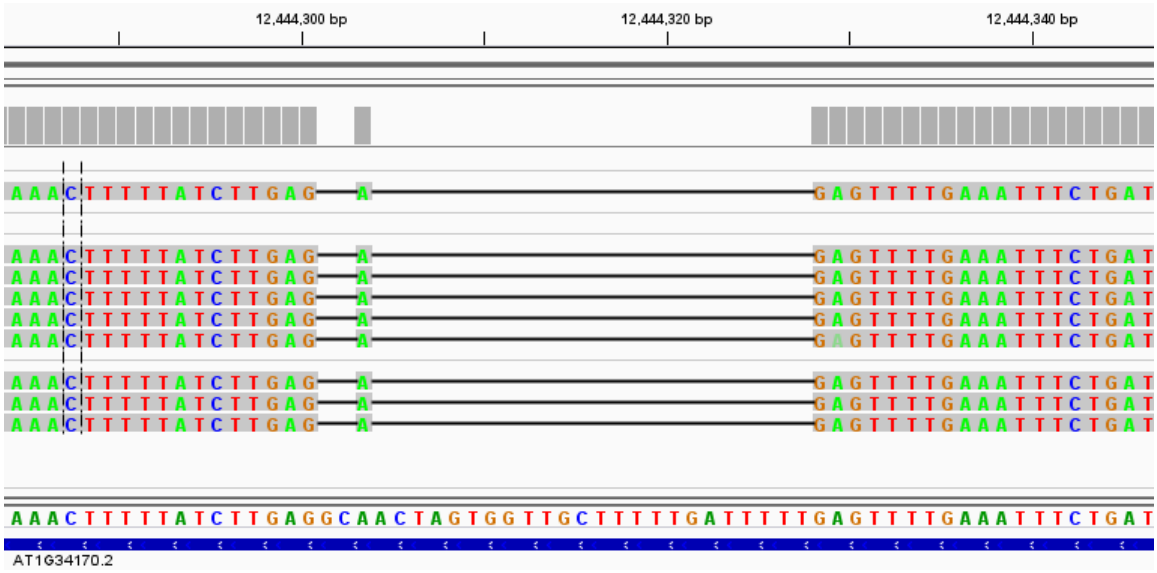
Although the purpose of using two different mapping programs, NGM and MASS, was to complement and strengthen the support for certain putative causal SNPs, in this case, the two software programs actually found different candidates. The leading candidates for *ursu*, *arf13* and *rpp8*, were identified by either NGM or MASS, respectively. Neither of these two genes were identified as potential candidates by both software packages. This

disagreement can most likely be attributed to differences in alignment and filtering criteria. These issues would most likely not arise in the case of higher sequence coverage, as including enough reads in the analyses would likely point to one (or only a few) causal candidates. Integrated Genomic Viewer (IGV) was used to view the sequence alignments used for NGM, and confirmed the presence of a large deletion in *arf13* (Figure 9). Searching the alignments used for NGM for the proposed location of the rpp8 SNP identified by MASS revealed no reads covering this region. This supports the idea that the difference in alignment programs used between NGM and MASS, compounded with dismal genomic coverage, may have contributed to the discovery of different putative causal mutations.

4.8 Alternative Modes of Trait Inheritance

The assumption has been made throughout this analysis that the root termination phenotype under study is a qualitative trait (i.e. individuals can be clearly identified according to phenotype, the trait shows Mendelian inheritance, and the trait shows few environmental effects). Prior to mapping by sequence analysis presented herein, 18 out of 200 plants were observed root terminators. This gives a ratio of ~ 0.09, versus the expected Mendelian ratio of $\frac{1}{16}$ or ~ 0.06. A χ^2 test comparing these two values reveals there is not a significant difference between observed and expected ratios of phenotypic variance in the F2 population (p-value ≥ 0.05). Observed segregation ratios and expected ratios of the phenotype were compared for the F2 population of the Col-0 x Ler cross, in which 64 out of 800 progeny were root terminators. A statistically significant difference was observed ($p \leq 0.05$). Differences in statistical significance can most likely be attributed to small sampling size; however, the larger sampling group suggests a

A.



B.

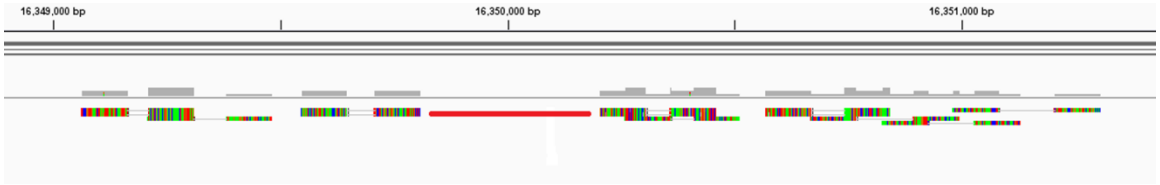


Figure 9. Illumina sequence coverage of selected genomic regions. Alignments were generated from the BAM file used for NGM analysis and viewed using Integrative Genomics Viewer (IGV). IGV confirmed low coverage across *rpp8* and *atr* regions. (A) Manual IGV alignment verified the presence of a major deletion in exon 12 of *arf13* (AT1G34170.2). (B) Exon 10 of *atr* was expected to possess a region without alignment, due to the T-DNA insertion. This region was confirmed (indicated by a red bar in the image above), although poor coverage created many such regions in *atr*.

deviation from expected Mendelian ratios. In this case, the phenotype could be reevaluated as a quantitative trait (i.e. there is a range of variation in phenotype, a complex, non-Mendelian mode of inheritance is shown, and there is moderate-high environmental effects). Examples of quantitative traits include plant height, seed weight, etc. A complex mode of inheritance may suggest the involvement of a quantitative trait locus (QTL). QTLs have been notoriously difficult to map and identify, although some progress using mapping by sequencing has been successful [99, 100]. The identity of this phenotype as a QTL may help to explain the micro-deserts seen using NGM.

Another possibility that has not been addressed is whether the phenotype is only dependent on the absence of functional ATR and not on an additional gene, such as *ursu*. Although permanent root arrest was not observed in UV-B irradiated *atr* mutants, the phenotype could be partially penetrant. The likelihood of this scenario seems low, however, because one would expect a much lower incidence of the phenotype. In order to account for sampling error and stochastic effects, additional *atr* mutants would need to be screened for the phenotype.

4.9 Conclusions

High-throughput sequencing technologies have transformed the field of genetics. Revolutionary mapping by screening techniques enable forward genetic screens because of affordability and efficiency. Paralleled progress of bioinformatics analysis programs continues to streamline the analysis of gargantuan datasets into informative, biologically relevant results. Here, it was hypothesized that permanent root arrest of $atr^{-/-}$ Arabidopsis plants in response to UV-B irradiation is dependent on another gene, *ursu.* Due to poor sequence coverage and potential linkage to *atr*, we could not identify a causative

mutation and the hypothesis could not be confirmed. However, several candidate mutations have been revealed, including *arf13* and *rpp8* using NGM and MASS respectively. Several potential scenarios exist to address the unexplained phenotype. Least likely, the phenotype is exclusively due to the lack of *atr*, and the phenotype is partially penetrant. However, in this situation, one would expect the observed and expected Mendelian ratios to show a substantial deviation, larger than that observed here. Segregation analysis with a larger sampling size would be necessary to test this possibility. Returning to the original hypothesis, it remains a possibility that the root termination phenotype is dependent upon another gene in addition to *atr*. In this case, the causal mutation could be located along chromosome four or the short arm of chromosome five. MASS analysis of this situation was difficult due to the compounding factors of low coverage and a low density of SNPs listed in both regions. NGM analysis did not pick out any salient candidates in these regions, but this could be attributed to poor coverage and differences in alignment programs used. The causal mutation could be located along one of the other three chromosomes (chromosomes one, two, or three). MASS did not show any obvious enrichments for Col-0 for any of these chromosomes, nor did NGM return any multi-megabase SNP deserts. *Arf13* was selected from among many micro-deserts in NGM for its physiological relevance. PCR and Sanger sequencing confirmed the P → L SNP found in NGM, and also found a 26-bp deletion. In the two most likely gene models, these changes caused a frameshift mutation, which extended the protein length in both models by 120 amino acids. It is possible that an *arf13 atr* mutant is experiencing loss of its auxin gradient and is unable to reestablish critical auxin gradients across the root tip, thereby resulting in the observed root termination

phenotype. A complementation test and a reverse genetics approach are necessary to validate these results. Finally, there is the possibility that the causal mutation is tightly linked to *atr*. This possibility would explain, despite low coverage, the presence of a salient peak and SNP desert output from both analyses. Low coverage may certainly explain the reason why another peak/desert was not observed, however it is interesting, despite poor coverage, that the positive control, *atr*, was so salient. Perhaps the situation is that low coverage is preventing greater resolution in this area. MASS proposed a list of candidates in this area, including a number of small RNA genes, as well as rpp8. NGM did not find these same candidates, likely due to differences in alignment algorithms. Finally, additional segregation analysis could be performed on the *ursu* line to better determine its inheritance pattern and test the nature of the phenotypic trait.

In order to provide more substantial support for this hypothesis and better test its predictions, greater genomic coverage – at least 25X – would be needed. A large population of mutant F2 plants for library preparation, as well as prior concentration of DNA and optimization of sonication time would be important steps for increasing DNA yields submitted for sequencing. Mapping programs NGM and MASS could be run again with more sequence data, and would likely return a more concise list of candidates. NIKS could also be utilized in a future analysis. The advantages to using NIKS are ease of applicability to mapping projects involving crops and other non-model organisms, and avoiding the need to cross mutants with a mapping line and collecting F2 individuals.

In conclusion, additional experimental verification is necessary to identify the gene responsible for the *ursu* phenotype. Evidence suggests the causal mutation may be linked to *atr*. *Rpp8* was herein identified as one potential candidate, although experimental

validation is still needed to assess the role of *rpp8* in causing root termination. Alternatively, a putative causal mutation has been identified here in *arf13*, which may suggest a link between signaling in the DNA damage response and essential phytohormone gradients.

Bibliography

1. Jackson, S.P. and J. Bartek, *The DNA-damage response in human biology and disease.* Nature, 2009. **461**(7267): p. 1071-1078.

2. Leonard, J.M., *Reduction of Stability of Arabidopsis Genomic and Transgenic DNA-Repeat Sequences (Microsatellites) by Inactivation of AtMSH2 Mismatch-Repair Function.* Plant Physiology, 2003. **133**(1): p. 328-338.

3. Modrich, P., *Mechanisms in Eukaryotic Mismatch Repair.* Journal of Biological Chemistry, 2006. **281**(41): p. 30305-30309.

4. Sancar, A., *DNA Excision Repair.* Annual Review of Biochemistry, 1996. **65**: p. 43-81.

5. Huang, J.-C., et al., *Substrate spectrum of human excinuclease: Repair of abasic sites, methylated bases, mismatches, and bulky adducts.* Proceedings of the National Academy of Sciences, 1994. **91**: p. 12213-12217.

6. Kairupan, C. and R.J. Scott, *Base excision repair and the role of MUTYH.* Hereditary Cancer in Clinical Practice, 2007. **5**(4): p. 199-209.

7. San Filippo, J., P. Sung, and H. Klein, *Mechanism of Eukaryotic Homologous Recombination.* Annual Review of Biochemistry, 2008. **77**(1): p. 229-257.

8. Lieber, M.R., *The Mechanism of Human Nonhomologous DNA End Joining.* Journal of Biological Chemistry, 2007. **283**(1): p. 1-5.

9. Kennedy, R.D., *The Fanconi Anemia/BRCA pathway: new faces in the crowd.* Genes & Development, 2005. **19**(24): p. 2925-2940.

10. Huen, M.S.Y. and J. Chen, *The DNA damage response pathways: at the crossroad of protein modifications.* Cell Research, 2008. **18**(1): p. 8-16.

11. Keightley, P.D. and M. Lynch, *Toward a realistic model of mutations affecting fitness.* Evolution, 2003. **57**(3): p. 683-685.

12. Ossowski, S., et al., *The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana.* Science, 2009. **327**(5961): p. 92-94.

13. Müller, A. and R. Fishel, *Mismatch Repair and the Hereditary Non-polyposis Coiorectal Cancer Syndrome (HNPCC).* Cancer Investigation, 2002. **20**(1): p. 102-109.

14. Yang, H., et al., *Identification of mutator genes and mutational pathways in Escherichia coli using a multicopy cloning approach.* Molecular Microbiology, 2004. **53**(1): p. 283-295.

15. Echols, H., C. Lu, and P.M.J. Burgers, *Mutator strains of Escherichia coli, mutD and dnaQ, with defective exonucleolytic editing by DNA polymerase III holoenzyme.* Proceedings of the National Academy of Sciences, 1983. **80**: p. 2189-2192.

16. Sikora, P., et al., *Mutagenesis as a Tool in Plant Genetics, Functional Genomics, and Breeding.* International Journal of Plant Genomics, 2011. **2011**: p. 1-13.

17. Rapoport, J.A., *Carbonyl compounds and the chemical mechanism of mutation.* Doklady Akademii Nauk, 1946. **54**: p. 65-67.

18. Shiwa, Y., et al., *Whole-Genome Profiling of a Novel Mutagenesis Technique Using Proofreading-Deficient DNA Polymerase δ.* International Journal of Evolutionary Biology, 2012. **2012**: p. 1-8.

19. Cimprich, K.A. and D. Cortez, *ATR: an essential regulator of genome integrity.* Nature Reviews Molecular Cell Biology, 2008. **9**(8): p. 616-627.

20. Bradbury, J.M. and S.P. Jackson, *ATM and ATR.* Cell, 2003. **13**(12): p. R468.

21. Yu, H., *Chk1: A Double Agent in Cell Cycle Checkpoints.* Developmental Cell, 2007. **12**(2): p. 167-168.

22. Stokes, M.P., et al., *Profiling of UV-induced ATM/ATR signaling pathways.* Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19855-19860.

23. Savitsky, K., et al., *A single ataxia telangiectasia gene with a product similar to PI-3 kinase.* Science, 1995. **23**: p. 1749-1753.

24. Wynbrandt, J. and M. Ludman, *Seckel syndrome*, in *The Encyclopedia of Genetic Disorders and Birth Defects*. 2009, Infobase Publishing: New York. p. 344.

25. Zou, L., *Sensing DNA Damage Through ATRIP Recognition of RPA-ssDNA Complexes.* Science, 2003. **300**(5625): p. 1542-1548.

26. Kumagai, A., et al., *TopBP1 Activates the ATR-ATRIP Complex.* Cell, 2006. **124**(5): p. 943-955.

27. Jiang, G. and A. Sancar, *Recruitment of DNA Damage Checkpoint Proteins to Damage in Transcribed and Nontranscribed Sequences.* Molecular and Cellular Biology, 2005. **26**(1): p. 39-49.

28. Rédei, G.P., *Arabidopsis as a Genetic Tool.* Annual Review of Genetics, 1975. **9**: p. 111-127.

29. Hays, J.B., *Arabidopsis thaliana, a versatile model system for study of eukaryotic genome-maintenance functions.* DNA Repair, 2002. **1**: p. 579-600.

30. Koncz, C. and G.P. Redei, *Genetic Studies with Arabidopsis: A Historical View*, in *Arabidopsis*, E.M. Meyerowitz and C.R. Somerville, Editors. 1994, Cold Spring Harbor Laboratory Press Plainview, NY. p. 223-240.

31. Meyerowitz, E.M., *Plants Compared to Animals: The Broadest Comparative Study of Development.* Science, 2002. **295**(5559): p. 1482-1485.

32. Bitonti, M.B. and A. Chiappetta, *Root Apical Meristem Pattern: Hormone Circuitry and Transcriptional Networks.* Progress in Botany, 2011. **72**: p. 37-71.

33. Schiefelbein, J.W. and P.N. Benfey, *Root Development in Arabidopsis*, in *Arabidopsis*, E.M. Meyerowitz and C.R. Somerville, Editors. 1994, Cold Spring Harbor Laboratory Press: Plainview, NY. p. 335-351.

34. Radtke, F., *Self-Renewal and Cancer of the Gut: Two Sides of a Coin.* Science, 2005. **307**(5717): p. 1904-1909.

35. Jones, D.L. and A.J. Wagers, *No place like home: anatomy and function of the stem cell niche.* Nature Reviews Molecular Cell Biology, 2008. **9**(1): p. 11-21.

36. Curtis, M.J. and J.B. Hays, *Cooperative responses of DNA-damage-activated protein kinases ATR and ATM and DNA translesion polymerases to replication-blocking DNA damage in a stem-cell niche.* DNA Repair, 2011. **10**(12): p. 1272-1281.

37. Britt, A.B., *Repair of DNA damage induced by solar UV.* Photosynthesis Research, 2004. **81**: p. 105-112.

38. Lehmann, A.R., D. McGibbon, and M. Stefanini, *Xeroderma pigmentosum.* Orphanet Journal of Rare Diseases, 2011. **6**(1): p. 70.

39. Yoon, J., et al., *The DNA Damage Spectrum Produced by Simulated Sunlight.* Journal of Molecular Biology, 2000. **299**: p. 681-693.

40. Furukawa, T., et al., *A shared DNA-damage-response pathway for induction of stem-cell death by UVB and by gamma irradiation.* DNA Repair, 2010. **9**(9): p. 940-948.

41. Yoshiyama, K., et al., *Suppressor of gamma response 1 (SOG1) encodes a putative transcription factor governing multiple responses to DNA damage.* Proceedings of the National Academy of Sciences, 2009. **106**(31): p. 12843-12848.

42. Koornneef, M., *Arabidopsis Genetics*, in *Arabidopsis*, E.M. Meyerowitz and C.R. Somerville, Editors. 1994, Cold Springs Harbor Laboratory Press: Plainview, NY. p. 89-116.

43. Blumenstiel, J.P., et al., *Identification of EMS-Induced Mutations in Drosophila melanogaster by Whole-Genome Sequencing.* Genetics, 2009. **182**(1): p. 25-32.

44. Kircher, M. and J. Kelso, *High-throughput DNA sequencing - concepts and limitations.* BioEssays, 2010. **32**(6): p. 524-536.

45. Soon, W.W., M. Hariharan, and M.P. Snyder, *High-throughput sequencing for biology and medicine.* Molecular Systems Biology, 2013. **9**.

46. Liu, L., et al., *Comparison of Next-Generation Sequencing Systems.* Journal of Biomedicine and Biotechnology, 2012. **2012**: p. 1-11.

47. Bentley, D.R., et al., *Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry.* Nature, 2008. **456**(7218): p. 53-59.

48. Cuperus, J.T., et al., *Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing.* Proceedings of the National Academy of Sciences, 2009. **107**(1): p. 466-471.

49. Austin, R.S., et al., *Next-generation mapping of Arabidopsis genes.* The Plant Journal, 2011. **67**(4): p. 715-725.

50. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-714.

51. Fahlgren, N., et al., *Computational and analytical framework for small RNA profiling by high-throughput sequencing.* RNA, 2009. **15**(5): p. 992-1002.

52. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Research, 2008. **18**(11): p. 1851-1858.

53. Schneeberger, K., et al., *SHOREmap: simultaneous mapping and mutation identification by deep sequencing.* Nature, 2009. **6**(8): p. 550-551.

54. Sarin, S., et al., *Caenorhabditis elegans mutant allele identification by whole-genome sequencing.* Nature Methods, 2008. **5**(10): p. 865-867.

55. Zuryn, S., et al., *A Strategy for Direct Mapping and Identification of Mutations by Whole-Genome Sequencing.* Genetics, 2010. **186**(1): p. 427-430.

56. Nordström, K.J.V., et al., *Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers.* Nature, 2013. **31**(4): p. 325-330.

57. Culligan, K., *ATR Regulates a G2-Phase Cell-Cycle Checkpoint in Arabidopsis thaliana.* The Plant Cell Online, 2004. **16**(5): p. 1091-1104.

58. Preuss, S.B. and A.B. Britt, *A DNA-Damage-Induced Cell Cycle Checkpoint in Arabidopsis.* Genetics, 2003. **164**: p. 323-334.

59. Curtis, M.J. and J.B. Hays, *Tolerance of dividing cells to replication stress in UVB-irradiated Arabidopsis roots: Requirements for DNA translesion polymerases η and ζ.* DNA Repair, 2007. **6**(9): p. 1341-1358.

60. Chang, C., et al., *Restriction fragment length polymorphism linkage map for Arabidopsis thaliana.* Proceedings of the National Academy of Sciences, 1988. **85**: p. 6856-6860.

61. King, G., J. Nienhuis, and C. Hussey, *Genetic similarity among ecotypes of Arabidopsis thaliana estimated by analysis of restriction fragment length polymorphisms.* Theoretical Applied Genetics, 1993. **86**: p. 1028-1032.

62. Williams, J.G.K., et al., *Genetic mapping of mutations using phenotypic pools and mapped RAPD markers.* Nucleic Acids Research, 1993. **21**: p. 2697-2702.

63. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler Transform.* Bioinformatics 2009. **25**(14): p. 1754-60.

64. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

65. Rozen, S. and H.J. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers*. 2000, Totowa, NJ: Humana Press.

66. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Research, 1994. **22**(22): p. 4673-4680.

67. Lamesch, P., et al., *The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.* Nucleic Acids Research, 2011. **40**(D1): p. D1202-D1210.

68. Greene, E.A., et al., *Spectrum of Chemically Induced Mutations From a Large-Scale Reverse-Genetic Screen in Arabidopsis.* Genetics, 2003. **164**: p. 731-740.

69. Schatz, M.C., A.L. Delcher, and S.L. Salzberg, *Assembly of large genomes using second-generation sequencing.* Genome Research, 2010. **20**(9): p. 1165-1173.

70. Srivatsan, A., et al., *High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies.* PLoS Genetics, 2008. **4**(8): p. e1000139.

71. Lister, R., B.D. Gregory, and J.R. Ecker, *Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond.* Current Opinion in Plant Biology, 2009. **12**(2): p. 107-118.

72. James, G., et al., *User guide for mapping-by-sequencing in Arabidopsis.* Genome Biology, 2013. **14**(6): p. R61.

73. Finnegan, E.J., *The small RNA world.* Journal of Cell Science, 2003. **116**(23): p. 4689-4693.

74. Winter, D., et al., *An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets.* PLoS ONE, 2007. **2**(8): p. e718.

75. Cooley, M.B., et al., *Members of the Arabidopsis HRT/RPP8 Family of Resistance Genes Confer Resistance to Both Viral and Oomycete Pathogens.* The Plant Cell, 2000. **12**: p. 663-676.

76. Kimchi-Sarfaty, C., et al., *A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity.* Science, 2007. **315**(5811): p. 525-528.

77. Grymek, K., et al., *Role of silent polymorphisms within the dopamine D1 receptor associated with schizophrenia on D1–D2 receptor hetero-dimerization.* Pharmacological Reports, 2009. **61**: p. 1024-1033.

78. O'Connell, M., et al., *In Arabidopsis thaliana codon volatility scores reflect GC3 composition rather than selective pressure.* BMC Research Notes, 2012. **5**(539): p. 1-12.

79. Saunders, R. and C.M. Deane, *Synonymous codon usage influences the local protein structure observed.* Nucleic Acids Research, 2010. **38**(19): p. 6719-6728.

80. Estelle, M. and H.J. Klee, *Auxin and Cytokinin in Arabidopsis*, in *Arabidopsis*, E.M. Meyerowitz and C.R. Somerville, Editors. 1994, Cold Springs Harbor Laboratory Press: Plainview, NY. p. 555-574.

81. Tiwari, S.B., *The Roles of Auxin Response Factor Domains in Auxin-Responsive Transcription.* The Plant Cell Online, 2003. **15**(2): p. 533-543.

82. Ulmasov, T., G. Hagen, and T.J. Guilfoyle, *Activation and repression of transcription by auxinresponse factors.* Proceedings of the National Academy of Sciences, 1999. **96**: p. 5844-5849.

83. Benjamins, R. and B. Scheres, *Auxin: The Looping Star in Plant Development.* Annual Review of Plant Biology, 2008. **59**(1): p. 443-465.

84. Guilfoyle, T.J. and G. Hagen, *Auxin response factors.* Current Opinion in Plant Biology, 2007. **10**(5): p. 453-460.

85. Kieffer, M., J. Neve, and S. Kepinski, *Defining auxin response contexts in plant development.* Current Opinion in Plant Biology, 2010. **13**(1): p. 12-20.

86. Petersson, S.V., et al., *An Auxin Gradient and Maximum in the Arabidopsis Root Apex Shown by High-Resolution Cell-Specific Analysis of IAA Distribution and Synthesis.* The Plant Cell Online, 2009. **21**(6): p. 1659-1668.

87. Paponov, I., et al., *The PIN auxin efflux facilitators: evolutionary and functional perspectives.* Trends in Plant Science, 2005. **10**(4): p. 170-177.

88. Petrasek, J. and J. Friml, *Auxin transport routes in plant development.* Development, 2009. **136**(16): p. 2675-2688.

89. Aida, M., et al., *The PLETHORA Genes Mediate Patterning of the Arabidopsis Root Stem Cell Niche.* Cell, 2004. **119**(1): p. 109-120.

90. Okushima, Y., *Functional Genomic Analysis of the AUXIN RESPONSE FACTOR Gene Family Members in Arabidopsis thaliana: Unique and Overlapping Functions of ARF7 and ARF19.* The Plant Cell Online, 2005. **17**(2): p. 444-463.

91. Guilfoyle, T.J. and G. Hagen, *Auxin Response Factors.* Journal of Plant Growth Regulation, 2001. **20**(3): p. 281-291.

92. Rademacher, E.H., et al., *A cellular expression map of the Arabidopsis AUXIN RESPONSE FACTOR gene family.* The Plant Journal, 2011. **68**(4): p. 597-606.

93. Weijers, D., et al., *Developmental specificity of auxin response by pairs of ARF and Aux/IAA transcriptional regulators.* The EMBO Journal, 2005. **24**: p. 1874-1885.

94. Shell, S.M., et al., *Checkpoint Kinase ATR Promotes Nucleotide Excision Repair of UV-induced DNA Damage via Physical Interaction with Xeroderma Pigmentosum Group A.* Journal of Biological Chemistry, 2009. **284**(36): p. 24213-24222.

95. Auclair, Y., et al., *ATR kinase is required for global genomic nucleotide excision repair exclusively during S phase in human cells.* Proceedings of the National Academy of Sciences, 2008. **105**(46): p. 17896-17901.

96. Vrouwe, M.G., et al., *UV-induced photolesions elicit ATR-kinase-dependent signaling in non-cycling cells through nucleotide excision repair-dependent and -independent pathways.* Journal of Cell Science, 2011. **124**(3): p. 435-446.

97. van den Berg, C., et al., *Short-range control of cell differentiation in the Arabidopsis root meristem.* Nature, 1997. **390**: p. 287- 289.

98. Sabatini, S., et al., *An Auxin-Dependent Distal Organizer of Pattern and Polarity in the Arabidopsis Root.* Cell, 1999. **99**: p. 463-472.

99. Meuwissen, T., *Use of whole genome sequence data for QTS mapping and genomic selection*, in *Proceeding of the ninth world congress on genetics applied*

to livestock production. 2010, Gesellschaft für Tierzuchtwissenschaften e. V: Leipzig, Germany.

100. Ober, U., et al., *Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster*. PLoS Genetics, 2012. **8**(5): p. e1002685.