
Somatic Mutations Drive Vascular Malformation Initiation, Progression, and Predisposition

by

Daniel Aaron Snellings

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

Douglas Marchuk, Supervisor

Beth Sullivan

Michael Hauser

Timothy Reddy

Craig Lowe

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Molecular Genetics and Microbiology
in the Graduate School of Duke University
2021

ABSTRACT

Somatic Mutations Drive Vascular Malformation Initiation, Progression, and Predisposition

by

Daniel Aaron Snellings

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

Douglas Marchuk, Supervisor

Beth Sullivan

Michael Hauser

Timothy Reddy

Craig Lowe

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Molecular Genetics and
Microbiology
in the Graduate School of Duke University
2021

Copyright © 2021 by Daniel Aaron Snellings
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Vascular malformations are a diverse class of focal lesions that may occur throughout the body and affect different vascular beds (e.g. arteries, veins, capillaries). Distinct from cancer, vascular malformations remain functional structures and lack the capacity for metastasis and uncontrolled growth. Despite their differences, vascular malformations and cancer share an underlying pathogenic mechanism: somatic mutations. The majority of somatic mutations in vascular malformations cause gain of function (GOF) in genes involved in vascular development and/or general proliferation (e.g. *TIE2*, *PIK3CA*, *KRAS*). Conversely, some vascular malformations such as CCM and CM-AVM are caused by loss of function (LOF) somatic mutations.

If you want to dedicate your thesis to anyone do so here

Contents

Abstract	iv
List of Tables	xii
List of Figures	xiii
List of Abbreviations and Symbols	xiv
Acknowledgements	xv
1 Introduction	1
1.1 Vascular Malformations	2
1.1.1 Common Genetic Mechanisms	2
1.2 Hereditary Hemorrhagic Telangiectasia	2
1.2.1 Genetics	2
1.2.2 Signaling of <i>ACVRL1</i> , <i>ENG</i> , and <i>SMAD4</i>	2
1.2.3 Relationship with Sporadic Arteriovenous Malformations	2
1.3 Sturge-Weber Syndrome	2
1.3.1 Mosaic Mutation of <i>GNAQ</i> p.R183Q	2
1.3.2 Function and Activity of GNAQ	2
1.3.3 Mutation of <i>GNAQ</i> in Other Diseases	2
1.4 Cerebral Cavernous Malformations	2
1.4.1 Genetics	2
1.4.2 Differences Between Familial and Sporadic Disease	2

1.4.3	Two-Hit Mechanism	2
1.4.4	Signaling of the CCM Complex and its Downstream Effectors	2
1.5	Infantile Hemangioma (??? Include Neg Data ???)	2
2	Two-Hit Mechanism of Hereditary Hemorrhagic Telangiectasia	3
2.1	Abstract	3
2.2	Introduction	4
2.3	Results	5
2.3.1	Telangiectasia Harmer a Somatic Mutation in <i>ENG</i> or <i>ACVRL1</i>	6
2.3.2	Somatic and Germline Mutations are Biallelic	10
2.3.3	Mutations are Consistent with Homozygous Loss of Function .	12
2.3.4	Telangiectasia from the Same Individual Harbor Unique Somatic Mutations	15
2.4	Discussion	17
2.4.1	Evidence for a Genetic Two-Hit Mechanism	17
2.4.2	Necessary, but Not Sufficient	20
2.4.3	Sensitivity for Detecting Somatic Mutations	20
2.4.4	Mutant Cell Metastasis	21
2.4.5	Probability of Multiple Somatic Mutations	22
2.4.6	Two-Hit Mechanism for <i>SMAD4</i> & JP-HHT	23
2.4.7	Compound Heterozygosity	23
2.4.8	Therapeutic Potential	24
2.5	Methods	24
2.6	Acknowledgements	31
3	Mutant <i>GNAQ</i> Alleles Produce Distinct Disease Phenotypes	32
3.1	Premise	32
3.2	Results	32

3.2.1	Port Wine Stain with Uncommon Mutation <i>GNAQ</i> p.Q209R	32
3.2.2	Structural Analysis of Common <i>GNAQ</i> Alleles	32
3.2.3	Functional Analysis of Common <i>GNAQ</i> Alleles (???Not sure it is appropriate to add this???)	32
3.2.4	Transcriptional Analysis of Common <i>GNAQ</i> Alleles	32
3.3	Discussion	32
3.3.1	Distinct Functions of <i>GNAQ</i> Alleles May Underlie Disease Specificity	32
3.3.2	Relationship Between <i>GNAQ</i> Activity and Mutation Timing . .	32
3.3.3	Importance of Mutated Cell Type in Determining Disease Fate	32
3.4	Methods	32
4	<i>MAP3K3</i> Mutations Seed Cerebral Cavernous Malformations	33
4.1	Premise	33
4.2	Results	33
4.2.1	<i>MAP3K3</i> Somatic Mutations Only Occur in Sporadic CCM . .	33
4.2.2	<i>MAP3K3</i> and CCM Gene Mutations are Mutually Exclusive .	33
4.2.3	Mutations in <i>KLF4</i> Do Not Contribute to CCM	33
4.2.4	(Whole-Exome Results) (??? Merge with above ???)	33
4.3	Discussion	33
4.3.1	CCM Loss of Function and <i>MAP3K3</i> Gain of Function are Functionally Equivalent	33
4.3.2	Differing Constraints for Constitutional and Somatic Inheritance	33
4.4	Methods	33
5	<i>PIK3CA</i> Mutations Fuel Cerebral Cavernous Malformation Growth	34
5.1	Premise	35
5.2	Results	35
5.2.1	<i>PIK3CA</i> Mutations Occur in Familial and Sporadic CCMs . .	35

5.2.2	CCMs Harbor Multiple Somatic Mutations in Different Genes	35
5.2.3	<i>PIK3CA</i> and CCM/ <i>MAP3K3</i> Mutations in the Same Cell . . .	35
5.2.4	Developmental Venous Anomalies Predispose to Malformation	35
5.3	Discussion	35
5.3.1	Three-Hit Model of CCM Pathogenesis	35
5.3.2	Similarities to the Genetic Mechanism of Cancer	35
5.3.3	Role of Clonal Expansion in Mutagenesis	35
5.3.4	Therapeutic Implications	35
5.3.5	Distinct Properties of <i>PIK3CA</i> vs. CCM/ <i>MAP3K3</i> Mutations	35
5.3.6	DVA Predispose to CCM and Other PI3K-Related Diseases .	35
5.4	Methods	35
6	Developmental Venous Anomalies Predispose to Malformation	36
6.1	Premise	36
6.2	Results	36
6.3	Discussion	36
6.4	Methods	36
7	Conclusions & Musings	37
7.1	HHT Pathogenesis	37
7.2	CCM Pathogenesis	37
7.2.1	Regrowth after Surgical Resection	37
7.2.2	CCM & Meningioma	37
7.3	Developmental Venous Anomalies as a Primer for Disease	41
7.3.1	Association with Sporadic CCM	41
7.3.2	Association with Other Diseases	41
7.3.3	Cowden Syndrome	41

7.3.4	Implications	41
7.4	Other Vascular Malformations	41
7.4.1	Classification of Vascular Malformations and Vascular Tumors	41
7.4.2	Sturge-Weber Syndrome and Somatic Mutations in <i>GNAQ</i> . .	41
7.4.3	The Curious Case of Infantile Hemangioma	41
7.5	The Molecular Basis of Genetic Dominance	44
7.5.1	Phenotypic Dominance ≠ Genetic Dominance	44
7.5.2	Knudsons Fingerprint	44
7.5.3	The Diverse Functional Effects of Genetic Mutations	44
7.6	The Intersection of Somatic Mutagenesis and Evolution	44
7.6.1	The Creation of New Alleles	44
7.6.2	The Relationship Between Mutability and Fitness Landscape .	44
7.6.3	Recurring Mutations & Convergent Evolution	44
7.6.4	Clonal Evolution of Somatic Mutants	44
7.6.5	Cancer	44
7.7	Somatic Mutations	44
7.7.1	The Role of Somatic Mutations in Aging	44
7.7.2	Constitutional Intolerance & Somatic Permissiveness	44
7.7.3	Somatic Reversion of Pathogenic Mutations	44
7.7.4	What is the Consequence of RNA Mutations?	44
7.8	Innovation in the Sequencing Era	44
7.8.1	Detection of Somatic Mutations	44
7.8.2	Single Cell Sequencing	44
7.8.3	Utility of Rare Disease Research in Mechanistic Discovery . .	44
7.8.4	Data Democratization & Individual Privacy	44

7.8.5 Growing Importance of Informatics in Biology	44
A Probability of Multiple Somatic Mutations	45
Bibliography	46
Biography	49

List of Tables

2.1	Mutations Identified in HHT Telangiectasia	8
2.2	Phase of Somatic and Germline Mutations in Telangiectasia	12
2.3	Predicted Consequences of Germline and Somatic Mutations	16

List of Figures

2.1	Low Frequency Somatic Mutations Detected in Telangiectasia.	9
2.2	Establishing Phase of Germline and Somatic Mutations	11
2.3	Mutations in <i>ACVRL1</i> Disrupt Splicing	14
2.4	Each Telangiectasia is Seeded by a Unique Somatic Mutation	18
7.1	<i>KLF4</i> Mutations in COSMIC.	39

List of Abbreviations and Symbols

Symbols

Put general notes about symbol usage in text here. Notice this text is double-spaced, as required.

- ✗ A blackboard bold X . Neat.
- ✗ A caligraphic X . Neat.
- ✗ A fraktur X . Neat.
- ✗ A boldface X .
- ✗ A sans-serif X . Bad notation.
- ✗ A roman X .

Abbreviations

Long lines in the `symbolist` environment are single spaced, like in the other front matter tables.

AR	Aqua Regia, also known as hydrochloric acid plus a splash of nitric acid.
SHORT	Notice the change in alignment caused by the label width between this list and the one above. Also notice that this multiline description is properly spaced.
OMFGTXTMSG4ME	Abbreviations/Symbols in the item are limited to about a quarter of the textwidth, so don't pack too much in there. You'll bust the margins and it looks really bad.

Acknowledgements

Thank anyone you like here. It's good practice to thank every granting agency that's given you money since you've been ABD, any other school you visited during your research, and any professional society that's funded your travel.

1

Introduction

1.1 Vascular Malformations

1.1.1 Common Genetic Mechanisms

1.2 Hereditary Hemorrhagic Telangiectasia

1.2.1 Genetics

1.2.2 Signaling of ACVRL1, ENG, and SMAD4

1.2.3 Relationship with Sporadic Arteriovenous Malformations

1.3 Sturge-Weber Syndrome

1.3.1 Mosaic Mutation of GNAQ p.R183Q

1.3.2 Function and Activity of GNAQ

1.3.3 Mutation of GNAQ in Other Diseases

1.4 Cerebral Cavernous Malformations

1.4.1 Genetics

1.4.2 Differences Between Familial and Sporadic Disease

1.4.3 Two-Hit Mechanism

1.4.4 Signaling of the CCM Complex and its Downstream Effectors

1.5 Infantile Hemangioma (??? Include Neg Data ???)

2

Two-Hit Mechanism of Hereditary Hemorrhagic Telangiectasia

2.1 Abstract

Hereditary Hemorrhagic Telangiectasia (HHT) is a Mendelian disease characterized by vascular malformations including visceral arteriovenous malformations and mucosal telangiectasia. HHT is caused by loss-of-function mutations in one of 3 genes; *ENG*, *ACVRL1* or *SMAD4* and is inherited as an autosomal dominant condition. Intriguingly, the constitutional mutation causing HHT is present throughout the body, yet the multiple vascular malformations in individuals with HHT occur focally, rather than manifesting as a systemic vascular defect. This disconnect between genotype and phenotype suggests that a local event is necessary for the development of vascular malformations. We investigated the hypothesis that local somatic mutations seed the formation HHT-related telangiectasia in a genetic two-hit mechanism. We identified low-frequency somatic mutations in 9/19 telangiectasia using next-generation sequencing. We established phase for 7 of 9 samples, which confirm that the germline and somatic mutations in all 7 samples exist in trans configuration;

consistent with a genetic two-hit mechanism. These combined data suggest that biallelic loss of *ENG* or *ACVRL1* may be a required event in the development of telangiectasia, and that rather than haploinsufficiency, vascular malformations in HHT are caused by a Knudsonian two-hit mechanism.

2.2 Introduction

Hereditary Hemorrhagic Telangiectasia (HHT) is a Mendelian disease characterized by the development of multiple focal vascular malformations consisting of arteriovenous malformations in visceral organs and telangiectasia in mucosal and cutaneous tissue. The genetic etiology of HHT has been established and is caused by mutations in *ENG*(?), *ACVRL1*(?), and rarely *SMAD4*(?); all of which follow an autosomal dominant inheritance pattern. Despite our understanding of the genetics of and downstream pathways involved in HHT, the molecular mechanisms that initiate HHT-related vascular malformation are poorly understood. Early studies of the functional consequences of HHT causal mutations established that these result in the loss of function of the gene product. These findings, in combination with autosomal dominant inheritance, led to the presumption that vascular malformations result from haploinsufficiency of the mutated gene product.(???) However, haploinsufficiency does not account for why HHT-related vascular malformations occur as strictly focal lesions, despite the systemic presence of the causal germline mutation. This disconnect between genotype and phenotype led to an alternative long-standing hypothesis that HHT-related vascular malformations result from a Knudsonian two-hit mechanism, where a local somatic mutation in the wild type allele of the affected gene seeds the formation of focal lesions.

The only published study which directly addresses the two-hit hypothesis attempted to determine whether Endoglin was present on the endothelial lining of arteriovenous malformations from an individual with HHT with a causal mutation

in the corresponding gene *ENG*.(?) Endoglin immunostaining was visible in the vessel lining, albeit at low levels. The presence of Endoglin in HHT-associated vascular malformations would contradict a two-hit mechanism, however complete loss of staining might not be predicted to occur, especially with the heterogeneous—and potentially mosaic—tissue of an arteriovenous malformation that may only contain a minority of cells that harbor the somatic mutation. Previous attempts to address this hypothesis at the DNA level have been hampered by the limitations of past sequencing technology. The advent of next-generation sequencing has drastically increased our sensitivity for detecting low-frequency somatic mutations. Somatic mutations have been identified in a diverse array of vascular malformations(???????) including recent evidence that sporadic arteriovenous malformations, not associated with HHT, harbor somatic activating mutations in *KRAS* or *MAP2K1*.(??) Notably, a genetic two-hit mechanism is known to contribute to Cerebral Cavernous Malformations (CCM)(???) and Capillary Malformation-Arteriovenous Malformation Syndrome (CM-AVM)(?); like HHT, both diseases are caused by autosomal dominant loss of function mutations. Here we demonstrate that HHT-related telangiectasia contain biallelic mutations in *ENG* or *ACVRL1*, resulting in homozygous loss of function; evidence in support of the long-standing hypothesis that telangiectasia pathogenesis follows a genetic two-hit mechanism.

2.3 Results

To determine whether a genetic two-hit mechanism underlies HHT pathogenesis, we tested three underlying expectations of the two-hit mechanism: 1) telangiectasia contain a somatic mutation in the same gene as a germline mutation which causes HHT, 2) the somatic and germline mutations are biallelic, and 3) both mutations result in loss-of-function.

2.3.1 Telangiectasia Harms or Somatic Mutation in ENG or ACVRL1

We used capture-based library preparations to sequence 19 telangiectasia for the three genes mutated in HHT (*ENG*, *ACVRL1*, and *SMAD4*) and 13 other vascular malformation-related genes (see Methods for identity of genes). The 13 non-HHT vascular malformation genes were chosen with the possibility that they also might harbor somatic mutations, but primarily to serve as control genes, since the two-hit mechanism requires a mutation in the corresponding HHT gene harboring the causal germline mutation. Somatic mutations in these other genes may or may not contribute to HHT pathogenesis, but absence of a somatic mutation in the casual HHT gene, would violate the first expectation of the genetic two-hit mechanism.

In each telangiectasia we identified a pathogenic germline mutation in either *ENG* or *ACVRL1*. Although in most cases the individuals' germline mutation was already known from clinical diagnostic sequencing, we intentionally remained blinded to this information until after our own sequence analysis of the tissue samples. In 6003-1, the individual harbors a silent germline mutation which was found by the clinical lab and noted as a variant of unknown significance. Below we show that this variant is indeed the pathogenic germline variant in this individual.

We used the MuTect2 variant caller to detect variants present in the sequence data. To identify candidate somatic mutations, we removed variants based on several stringent filtering criteria including briefly; intronic or intergenic variants, population frequency $>0.01\%$, $<0.1\%$ or <5 total supporting reads, low coverage, strand specificity, and low base quality scores. To validate or refute the authenticity of each candidate somatic mutation we performed an independent round of amplification using primers flanking each putative variant position for each sample, and sequenced to $>10000x$ coverage. In each tissue sample, the identical somatic variant was re-identified. Thus, these variants were bona fide somatic mutations

existing in the telangiectatic tissue. In total, we identified somatic variants in 9 of 19 telangiectasia; 5 in *ENG*(NM_001114753.1 (*ENG_v001*)) 4 in *ACVRL1*(NM_000020.2 (*ACVRL1_v001*)) (Table 2.1) (Figure 2.1) (See Methods). In each case, the somatic mutation was found in the same gene as the pathogenic germline mutation. Somatic mutations were not found in any of the other 15 genes sequenced, not even in one of the other HHT casual genes. Importantly, no telangiectasia harbored more than a single somatic mutation. The lack of mutational noise suggests these mutations are pathobiologically significant. Importantly, all are consistent with strong mutations; five of the variants are small indels resulting in a frameshift, three are in-frame indels, and one is a point mutation 4 bases after an exon-intron boundary that is predicted to impact RNA splicing.

Although these variants fall well below the 50% allele frequency expected for germline variants, it is formally possible that these variants exist constitutionally as very rare, somatic mosaic variation in the individual. We next investigated whether these somatic mutations were present in constitutional DNA from the individual. A source of constitutional DNA (saliva) was available for three of the nine mutation-positive samples, and in each we find that the somatic mutation was completely absent or present at a level no higher than technical sequencing noise in that sample (see Methods). Saliva was not available for 6001, but we obtained and sequenced DNA from multiple telangiectasia collected from this same individual. This enabled us to determine whether any of the five somatic mutations that we identified in individual samples was present in tissue of near identical pathobiology from the same individual; compared with saliva as a control, this is a more powerful test for somatic mosaicism. We found that the somatic mutations identified in five of the telangiectasia for which we identified a mutation were entirely absent in all other telangiectasia from this same individual. Finally, Sample 6005-1 is a single archived FFPE telangiectasia and no source of constitutional DNA is available. In total, we

Table 2.1: Summary of Somatic Mutation Discovery and Validation Sequencing of HHT Telangiectasia.

Sample	Germline Mutation	Somatic Mutation	Discovery Reads ^a	Validation Reads ^a	Constitutional Reads ^a
6001-1	<i>ENG</i> c.1080_1083del	<i>ENG</i> c.293_304del	33/1318 (2.5%)	1067/100268 (1.1%)	0/26462 (0%)
6001-3	same as above	<i>ENG</i> c.1195_1196delAGfsX2	5/1080 (0.46%)	723/115963 (0.62%)	0/24357 (0%)
6001-7	same as above	<i>ENG</i> c.1237_1238insCAfsX7	27/5127 (0.53%)	341/115570 (0.30%)	0/23066 (0%)
6001-8	same as above	<i>ENG</i> c.578delCinsTGC p.T193MR	111/4845 (2.3%)	1142/142572 (0.80%)	0/21315 (0%)
6001-10	same as above	<i>ENG</i> c.205delGfsX6	33/3389 (1.0%)	3575/326894 (1.1%)	0/22098 (0%) ^b
6001-*	same as above	NF			
6002-1	<i>ACVRL1</i> c.1451G>A p.R484Q	<i>ACVRL1</i> c.349delGinsTTfsX52	20/2217 (0.90%)	309/24018 (1.3%)	0/65818 (0%)
6002-2	same as above	<i>ACVRL1</i> c.1378_3_1402del19ins9	26/1649 (1.6%)	3189/202550 (1.6%)	6/155855 (0.0038%)
6003-1	<i>ACVRL1</i> c.474A>T p.G158G	<i>ACVRL1</i> c.625+4A>T	101/3392 (3.0%)	372/16303 (2.3%)	2/38924 (0.0051%)
6004-1	<i>ACVRL1</i> c.1232G>A p.R411Q	NF			
6004-2	same as above	NF			
6005-1	<i>ACVRL1</i> c.1232G>A p.R411Q	<i>ACVRL1</i> c.1206delCfsX12	133/1664 (8.0%)	2671/189690 (1.4%)	N/A

*Eight additional telangiectasia from patient 6001 with no identified somatic mutation
For multiple telangiectasia collected from one individual the sample ID is listed as (Patient#)-(Telangiectasia#), NF = None Found

^aAllele frequency in other telangiectasia from 6001

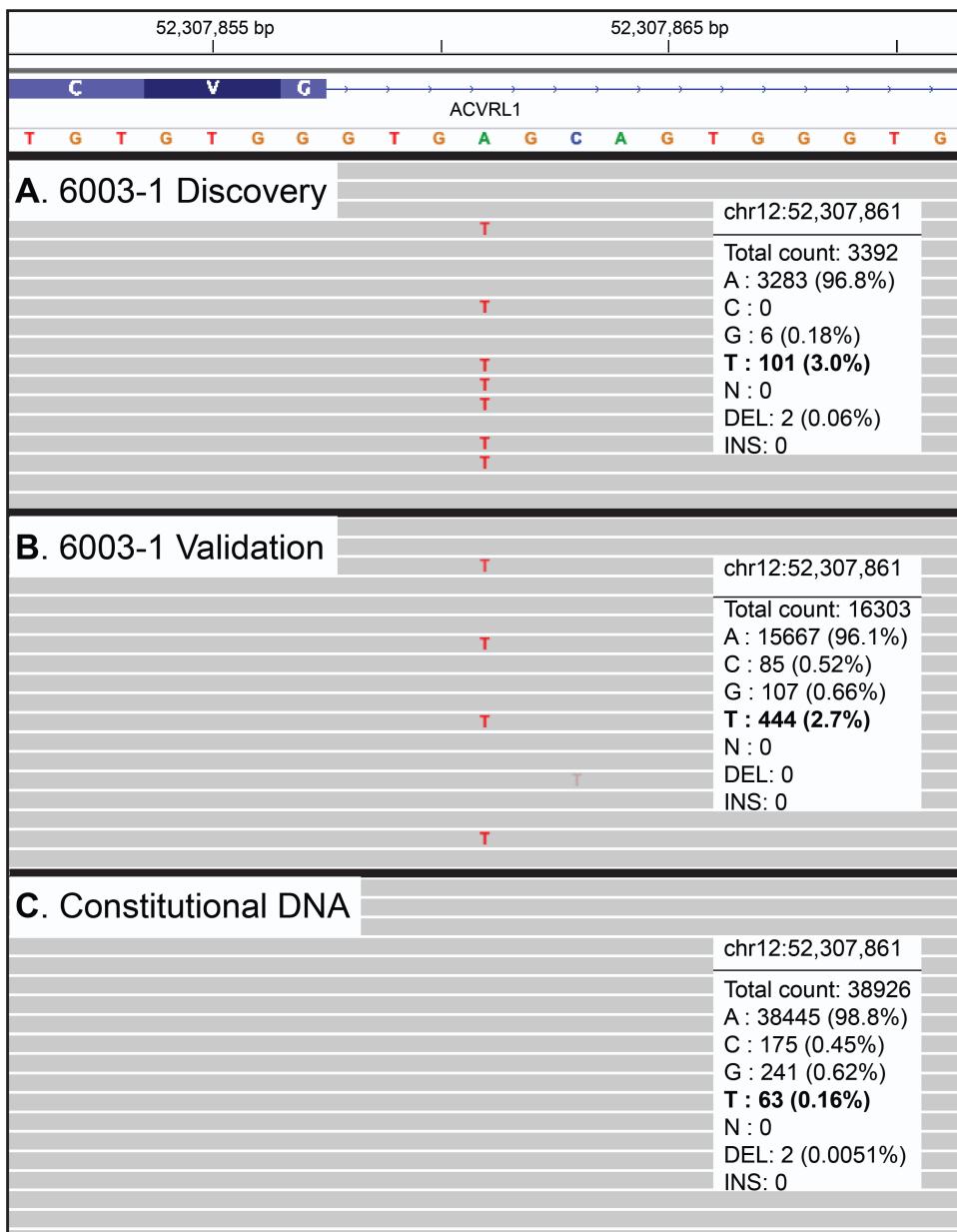


FIGURE 2.1: Visualization with IGV of next-generation sequencing data for one representative sample with a somatic mutation in *ACVR1L*. The somatic mutation in 6003-1, an A>T point mutation, is present at low frequency in DNA from telangiectatic tissue used for (A) discovery (capture-based sequencing) and (B) validation (amplicon-based sequencing). The mutation is below the level of sequencing noise in (C) constitutional DNA (amplicon-based sequencing) confirming the mutation is somatic.

found that 9 of 19 telangiectasia harbor a somatic mutation specifically in the same gene as a pathogenic germline mutation and that these mutations are not present constitutionally. The presence of somatic mutations in telangiectasia fulfills the 1st expectation of the genetic two-hit mechanism.

2.3.2 Somatic and Germline Mutations are Biallelic

The 2nd expectation of the genetic two-hit mechanism is that the somatic and germline mutations are biallelic; such that the somatic mutation occurs on the wild-type allele of the affected gene, in trans with the germline mutation. To determine if the mutations are biallelic we examined whether they were arranged in a cis or trans configuration by sequencing amplicons that cover the nucleotide positions of both somatic and germline mutations in a single molecule. The amplicons were sequenced with either short-read (Illumina) or long-read (PacBio) chemistry, depending on the amplicon size, in order to generate reads that would span the two mutations. In contrast to traditional Sanger sequencing which measures the population average at each position, both Illumina and PacBio chemistries output sequences of single DNA molecules. In total we generated mutation-spanning reads for 7 telangiectasia, each with more than 100 reads that contained the somatic mutation. From these mutation-spanning reads we established that >95% of reads with the somatic mutation possessed the wild type allele at the position of the germline mutation, showing that all 7 mutation pairs are in trans configuration (Table 2.2) (Figure 2.2A-B). Any two variants in a chromosome must be arranged in cis or trans with an equal probability of either arrangement. Considering this, our observation that 7/7 mutation pairs are arranged in trans corresponds to a p-value of 0.008 demonstrating significant bias towards a trans configuration. These data show that the somatic and germline mutations are biallelic, fulfilling the 2nd expectation of the genetic two-hit mechanism.

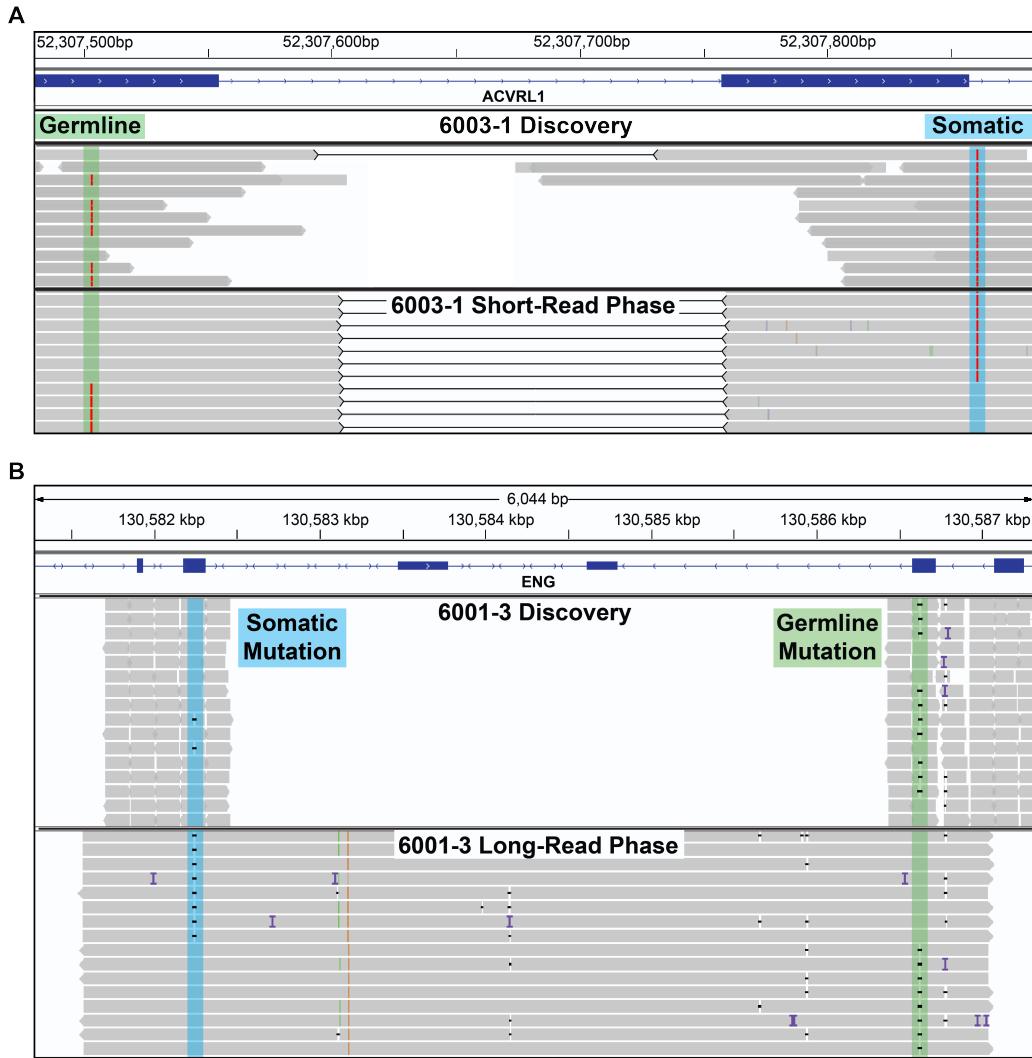


FIGURE 2.2: IGV visualization of two samples showing both methods of establishing phase. Each panel shows reads from the initial discovery sequencing and the reads used to establish phase. (A) Somatic and germline mutations in 6003-1 are both A>T point mutations and highlighted by the blue and green regions respectively. Since the distance between these mutations is relatively small (357bp) phase was established using Illumina short-reads (also used for validation). Black lines between reads denote read pairs, showing that both reads originate from a single molecule of DNA. Each molecule with the somatic mutation contains the wild-type allele at the germline mutation position proving the mutations are biallelic. (B) Somatic and germline mutations in 6001-3, both small deletions. The genomic distance between these mutations is 4377bp. In long reads that span the two mutations, each read with the somatic mutation contains the wild-type allele at the position of the germline mutation.

Table 2.2: Phase of Somatic and Germline Mutation Pairs.

Sample	Total Reads	Trans Reads	Cis Reads	P-value
6001-1	N/A	N/A	N/A	N/A
6001-3	112	112 (100%) ^a	0 (0%)	1.9e-34
6001-7	155	153 (98.7%) ^a	2 (1.3%)	2.6e-43
6001-8	593	590 (99.5%) ^a	3 (0.5%)	1.0e-171
6001-10	N/A	N/A	N/A	N/A
6002-1	125	120 (96.0%) ^a	5 (4.0%)	5.5e-30
6002-2	3189	3160 (99.0%) ^{bc}	29 (1.0%)	4.2e-890
6003-1	372	364 (97.8%) ^{bc}	6 (1.4%)	1.4e-99
6005-1	2671	2653 (99.3%) ^{bc}	18 (0.7%)	6.3e-759

^aReads generated with PacBio long-read chemistry)

^bReads generated with Illumina short-read chemistry

^cThese reads were also used for validation shown in Table 2.1

2.3.3 Mutations are Consistent with Homozygous Loss of Function

The 3rd expectation of the genetic two-hit mechanism is that the biallelic somatic and germline variants both result in loss of function. Due to the functional studies and extensive allelic series of mutations in each of the HHT genes, HHT is known to be caused by loss of function mutations. The germline mutation in 4 of the 5 individuals in this study has been identified previously in an individual with HHT and are reported in ClinVar (6001:VCV000213214.2, 6002:VCV000212796.2, 6004/6005:VCV000008243.2). There are also several publications supporting the pathogenicity of these mutations.(?????) These are all therefore bona fide loss of function mutations. The germline mutation in 6003-1, a silent mutation in *ACVRL1* exon 4, has been identified before in an individual with HHT, however it was classified as a variant of unknown significance (VUS). We used the in silico tool Human Splicing Finder 3.1(?) to analyze this variant and found that it was predicted to both

disrupt an exonic splice enhancer and create an internal splice donor site, potentially activating a cryptic splice site. Based on this prediction, we extracted RNA from peripheral blood leukocytes of 6003 and a control individual and performed RT-PCR to examine the splicing of *ACVRL1* transcripts. RNA from 6003 shows a new splice variant that is not present in control wild-type RNA (Figure 2.3B). As predicted by the Splice Finder program, the aberrant transcript is spliced precisely at the internal splice donor created by the mutation. The resulting transcript is missing the portion of exon 4 downstream of the germline mutation and skips exon 5 resulting in the in-frame deletion of 52 amino acids (Figure 2.3D-E). This deleted region contains several codons with known pathogenic missense mutations, suggesting that the 52 amino acid deletion would likely also result in loss of function. It is possible that the skipping of exon 5 is due to alternative splicing observed only in peripheral blood leukocytes, rather than a result of the mutation. If exon 5 is retained, the mutation would then generate a protein lacking 17 amino acid residues from exon 4 but then be frameshifted for the remainder of the transcript. With this evidence, all of the identified germline mutations meet the American College of Medical Genetics (ACMG) criteria for pathogenic mutations(?), fulfilling the first half of the 3rd expectation of the genetic two-hit mechanism.

In contrast to the germline mutation, many of which have been previously identified, the somatic mutations we identified are all novel. The ACMG guidelines for establishing pathogenicity are not applicable to somatic variants, however several lines of evidence support that all of the somatic mutations result in loss of function. Five of the nine somatic mutations result in a frameshift; all resulting in premature termination codons which would generate transcripts susceptible to nonsense-mediated decay. Frameshift mutations in *ENG* or *ACVRL1* are the most common mechanism for loss of function leading to HHT. Based on this, the 5 somatic frameshift mutations likely result in loss of function. Other than frameshift

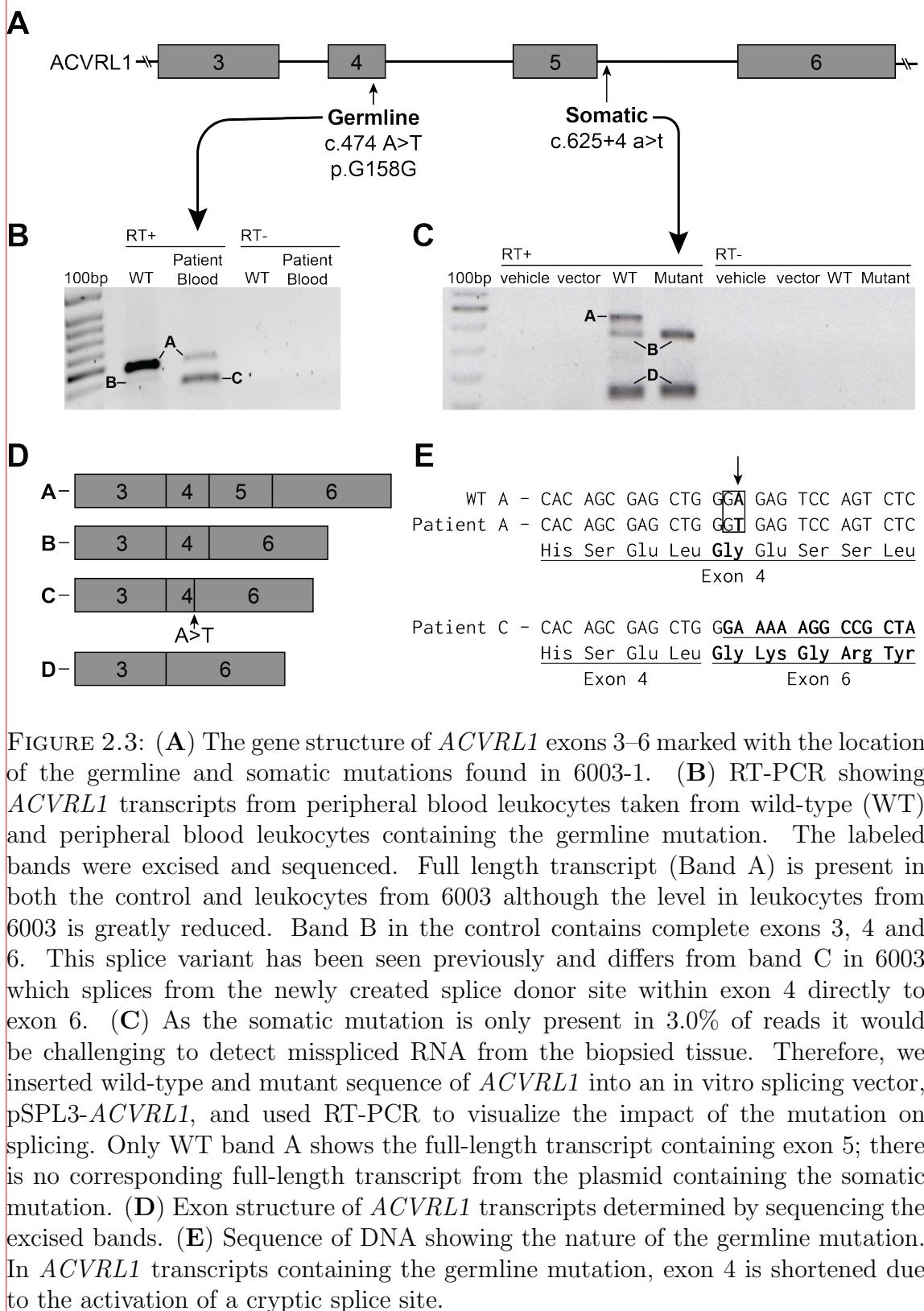


FIGURE 2.3: (A) The gene structure of *ACVRL1* exons 3–6 marked with the location of the germline and somatic mutations found in 6003-1. (B) RT-PCR showing *ACVRL1* transcripts from peripheral blood leukocytes taken from wild-type (WT) and peripheral blood leukocytes containing the germline mutation. The labeled bands were excised and sequenced. Full length transcript (Band A) is present in both the control and leukocytes from 6003 although the level in leukocytes from 6003 is greatly reduced. Band B in the control contains complete exons 3, 4 and 6. This splice variant has been seen previously and differs from band C in 6003 which splices from the newly created splice donor site within exon 4 directly to exon 6. (C) As the somatic mutation is only present in 3.0% of reads it would be challenging to detect misspliced RNA from the biopsied tissue. Therefore, we inserted wild-type and mutant sequence of *ACVRL1* into an in vitro splicing vector, pSPL3-*ACVRL1*, and used RT-PCR to visualize the impact of the mutation on splicing. Only WT band A shows the full-length transcript containing exon 5; there is no corresponding full-length transcript from the plasmid containing the somatic mutation. (D) Exon structure of *ACVRL1* transcripts determined by sequencing the excised bands. (E) Sequence of DNA showing the nature of the germline mutation. In *ACVRL1* transcripts containing the germline mutation, exon 4 is shortened due to the activation of a cryptic splice site.

mutations, the other 4 somatic mutations we identified consisted of 3 in-frame deletions and 1 intronic mutation predicted to impact splicing. These 4 mutations are not present in the genome aggregation database (gnomAD) showing that the population allele frequency of these variants is extremely low or zero. For the somatic in-frame deletion mutation found in 6001-8, there are two reports of different in-frame deletions with overlap at this position which are known to cause HHT, suggesting that the somatic deletion in 6001-8 is likely to result in loss of function. The somatic mutations in 6001-1 and 6002-2 also result in in-frame deletions, which delete 4 and 7 amino acids respectively. Comparing the crystal structures of *ENG* and *ACVRL1* we determined that the somatic mutations in 6001-1 and 6002-2 delete portions of a beta strand and helix respectively, potentially impacting protein folding (Table 2.3). The in silico tool PROVEAN was used to predict how the protein would tolerate these deletions. The threshold of -2.5 or lower (more negative) is considered a deleterious change. The scores for 6001-8, 6001-1, and 6002-2 were -6.106, -14.903, and -25.903 respectively, strongly suggesting that all three are deleterious. The remaining somatic mutation is intronic, and occurs 4 nucleotides from the exon-intron boundary. We used Human Splicing Finder 3.1 to predict the effect of this variant on splicing, and found that it is likely to disrupt the donor site. This prediction was confirmed by RT-PCR using an in vitro splicing construct which revealed that the somatic mutation prevents the formation of full-length *ACVRL1* transcripts (Figure 2.3C). In summary we present evidence supporting that the biallelic germline and somatic mutations all likely result in loss of function, fulfilling the 3rd expectation of the genetic two-hit mechanism.

2.3.4 Telangiectasia from the Same Individual Harbor Unique Somatic Mutations

We next sought to determine whether mutant cells in different telangiectasia derive from a somatic mutation in a common ancestor cell, or whether the mutant cell

Table 2.3: Evidence Germline and Somatic Mutations Result in Loss of Function.

Sample	Germline Mutation	Somatic Mutation
6001-1	Frameshift: PVS1 6 supporting publications	In-Frame Deletion (-4 residues) PROVEAN: Deleterious (-14.903) deletes region in β -sheet gnomAD AF: 0
6001-3	same as above	Frameshift common ENG LOF mechanism, expect NMD
6001-7	same as above	Frameshift common ENG LOF mechanism, expect NMD
6001-8	same as above	In-Frame Delins (-1 +2 residues) 2 pathogenic in-frame indels overlapping this codon ?? PROVEAN: Deleterious (-6.106) gnomAD AF: 0
6001-10	same as above	Frameshift common ENG LOF mechanism, expect NMD
6002-2	same as above	In-Frame Delins (-7 +1 residues) PROVEAN: Deleterious (-25.903) deletes region in α -helix gnomAD AF: 0
6003-1	Cryptic Splice Site: PS3 in silico predicted to activate cryptic site in vitro evidence (Figure 2.3)	Splice Site in silico predicted to disrupt donor site gnomAD AF: 0
6005-1	Missense: PS1 16 supporting publications	Frameshift common ACVRL1 LOF mechanism, expect NMD

AF = Allele Frequency; NMD = Nonsense Mediated Decay

Germline variant classification according to ACMG guidelines, not applicable to somatic variants(?)

PVS1 = Very strong evidence for pathogenicity; PS1-4 = Strong evidence

PROVEAN scores below -2.5 are predicted deleterious

population in each telangiectasia derives from an independent somatic mutation event. To test this, we examined the somatic mutations present in multiple telangiectasia from single individuals. In 6001, for which we had obtained 13 different telangiectasia, we identified a somatic mutation in 5 telangiectasia tissue samples. In each case the somatic mutation in each telangiectasia was unique. Likewise, for 6002, for which we had two telangiectasia, we identified a unique somatic mutation in each (Figure 2.4). These results are consistent with independent mutation events rather than the somatic mutation occurring in a progenitor cell or clonality due to a metastasis from a single initial lesion.

2.4 Discussion

2.4.1 Evidence for a Genetic Two-Hit Mechanism

In this study we present strong evidence that vascular malformations associated with HHT, specifically, cutaneous telangiectasia, follow a genetic two-hit mechanism of pathogenesis. HHT is also associated with arteriovenous malformations in lung, liver, brain and the gastrointestinal tract, but these tissues are not available for prospective collection. We postulate that the visceral, deeper vascular malformations that occur in HHT also follow this two-hit mechanism.

The two-hit hypothesis for HHT pathogenesis has persisted for decades without evidence, but these low frequency somatic mutations are challenging to identify using traditional sequencing methods. The only published study to address this topic employed immunohistochemical staining in an attempt to identify endothelial cells lacking staining in the lining of HHT-related arteriovenous malformations.(?) However, the absence of staining as a proxy for the gain of a mutation could be difficult to discover, especially if only a fraction of a cells would exhibit this lack of signal.

Using next-generation sequencing with unique molecular identifiers we success-

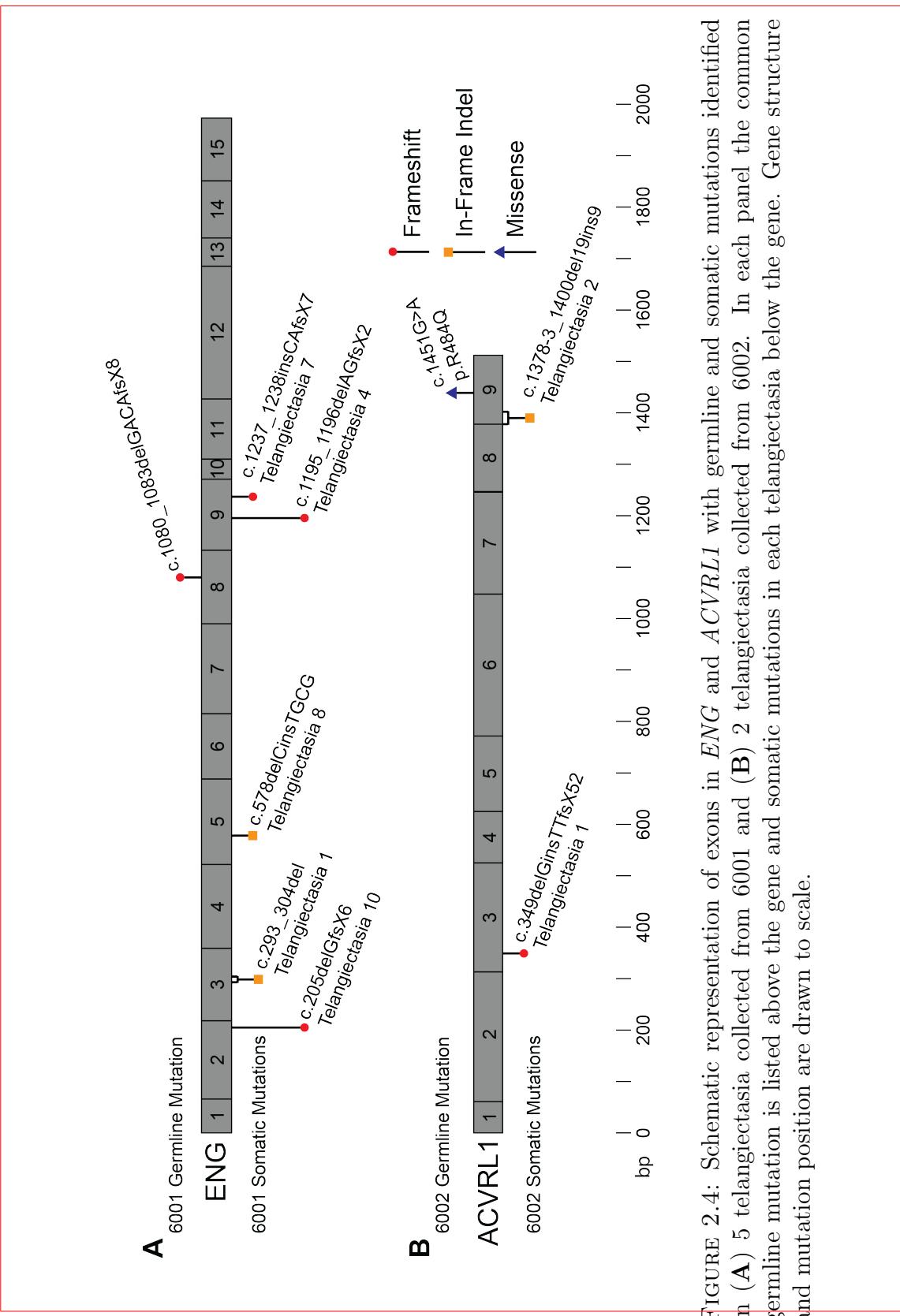


FIGURE 2.4: Schematic representation of exons in *EN_G* and *ACVRL1* with germline and somatic mutations identified in (A) 5 telangiectasia collected from 6001 and (B) 2 telangiectasia collected from 6002. In each panel the common germline mutation is listed above the gene and somatic mutations in each telangiectasia below the gene. Gene structure and mutation position are drawn to scale.

fully identified somatic mutations in multiple telangiectasia from different individuals with HHT. The somatic mutations we identified were present at frequencies ranging between 0.46% and 8.0% in the tissue, with an average of 2.3%. The low allele frequency is likely a result of two main contributing factors; the presence of normal tissue in the skin biopsy, and somatic mosaicism within the telangiectasia. The telangiectasias in this study were sampled as skin punch biopsies and although some of the surrounding skin tissue was removed before DNA extraction, an undetermined amount of normal tissue invariably remained. However, after removing the surrounding tissue, the enrichment for the vascular component of the tissue was subjectively greater than the often low somatic mutation allele frequency might suggest. We posit that a second explanation for the low mutant allele frequency is that telangiectasia are mosaic for the somatic mutation. This agrees with existing data from mouse models of HHT showing that induced retinal AVMs are mosaic: consisting of both heterozygous and homozygous null cells.(?) It is also consistent with the heterogeneity seen in Cerebral Cavernous Malformations(??) and the low mutant allele frequency in somatic mutations of other vascular malformation disorders.(????????????) Vascular malformations, similar to tumors in cancer, appear to be seeded by somatic mutations; however unlike tumors, vascular malformations do not appear to consist of pure populations of clonally expanded mutant cells, but contain a substantial percentage of unmutated cells.

In addition to the data presented here, the two-hit hypothesis of HHT-related vascular malformations is consistent with observations from mouse models of the disease. Whereas constitutional loss of both copies of *Eng* or *Acvrl1* in mice is embryonic lethal, mice heterozygous for constitutional deletion of either gene show extremely mild phenotypes with relatively few, if any, detectable vascular malformations.(??) Robust mouse models of HHT that recapitulate vascular malformation phenotypes require the use of Cre-Lox technology to delete both copies of *ENG* in a

temporally controlled (postnatal), cell-type specific (endothelial cells) manner.

In these mouse models of HHT it is also required that biallelic KO of the HHT gene occur in endothelial cells. Several groups have experimented with expressing Cre recombinase in different vascular-related cell types including pericytes (NG2-Cre), vascular smooth muscle cells (Myh11-Cre), and endothelial cells (Scl-Cre & Pdgfb-Cre); however, mice only develop vascular malformations when Cre is expressed in endothelial cells.(????) We have attempted to confirm that the somatic mutations we identified in human lesions occur in the endothelium by using laser capture microdissection, however these efforts have been hampered by the small quantity of tissue in telangiectasia biopsies and the difficulty of isolating a single layer of cells by microdissection. This question may be more easily addressed in larger arteriovenous malformations; however, these samples have been thus far inaccessible due to the rarity of their removal in individuals with HHT.

2.4.2 Necessary, but Not Sufficient

Interestingly, in addition to the local requirement for loss of both alleles, vascular malformations in this model only develop after injury such as an ear punch, or by VEGF injection.(?) This requirement of an angiogenic stimulus is consistent in mouse models for all HHT genotypes: *Eng*, *Acvrl1*, and *Smad4*.(?) The requirement for knockout of both copies of the gene supports the genetic two-hit mechanism we describe here. In addition, the necessity for an angiogenic stimulus suggests that loss of both copies of the relevant HHT gene is necessary, but not sufficient, for the development of the vascular malformation.

2.4.3 Sensitivity for Detecting Somatic Mutations

We might have expected to find somatic mutations in every telangiectasia, however we only found somatic mutations in 9 of the 19 we sequenced. The next-generation

sequencing strategy we employ for discovering somatic mutations is extremely sensitive for the detection of point mutations and small indels. However, there are several other types of genetic alterations that would result in biallelic loss of function due to loss of heterozygosity (LOH). LOH is a common occurrence in many tumors in cancer, and is a predominant mechanism of somatic loss/mutation. LOH can occur due to a variety of genetic mechanisms: large deletions, chromosome loss, and mitotic recombination. Given the apparent capacity for even a low fraction of somatically mutant cells to initiate the vascular malformation, it follows that the same would be true for LOH-associated mutational events; therefore, these mutations would appear instead as allelic imbalance rather than outright LOH. But if the level of allelic imbalance is as low as the frequency of somatic mutations we have observed in this study, we might expect linked marker haplotype ratios in the range of 48% to 52% at nearby markers; this slight and even trivial imbalance would be difficult if not impossible to detect and validate. It is also possible that non-genetic mechanisms such as loss of expression due to epigenetic silencing account for biallelic loss of function. This process, like the LOH associated events, would not be detected by our sequencing strategy; a problem that is only exacerbated by low allele frequency. Thus, it may not be surprising that we identified a somatic mutation in approximately only 9 of the VMs that were sequenced. We postulate it is highly likely that all of the 10 telangiectasia with no identified somatic mutation have biallelic loss by one of these other mechanisms.

2.4.4 Mutant Cell Metastasis

One consequence of the genetic two-hit mechanism that might appear to be improbable is that, if true, a new somatic mutation must occur in every one of the numerous vascular malformations in HHT. For example, some affected individuals have dozens or more visible telangiectasia on the skin and mucocutaneous surfaces alone.(???) An

attractive hypothesis to reconcile this conundrum would be if a somatic mutation first occurs in a circulating progenitor cell which then proliferates and seeds the formation of multiple telangiectasia. There is precedence for this mechanism in another vascular malformation syndrome, Blue Rubber Bleb Nevus syndrome. These individuals display multiple small vascular lesions which harbor an identical somatic double-mutation in the *TEK* gene. These vascular lesions appear to be anatomically-dispersed clones arising from an original, dominant, large lesion.(?) In HHT-related vascular malformations, we report evidence that contradicts this hypothesis: different telangiectasia collected from the same individual harbor different, unique somatic mutations. This observation does not exclude the possibility that circulating cells may in some cases spread telangiectasia, however the data thus far suggests that the primary mechanism is independent somatic mutation events.

2.4.5 Probability of Multiple Somatic Mutations

The dilemma of the requirement for numerous independent somatic mutation events in a single gene can be resolved with a probabilistic argument. Considering the size of the human genome (3.23×10^9 bp), the probability that a random somatic mutation occurs in the coding sequence of *ENG* (3201bp) in trans (50% likelihood) with a pathogenic *ENG* germline mutation is 0.00005%. Compounding this value with empirical evidence that single cells have anywhere from 100 to 1500 somatic mutations per cell(???), and an estimate that 5.66% of exonic somatic mutations result in LOF(?); we calculate a conservative estimate that 0.00028% of cell have biallelic LOF *ENG* mutations. An adult human has at least 6×10^{11} endothelial cells(?), therefore we estimate that an individual with HHT and a germline mutation in *ENG* has biallelic LOF *ENG* mutations in 1.5 million endothelial cells. It is clear that each cell with *ENG* biallelic LOF does not result in vascular malformation as individuals with HHT have at most hundreds, not millions, of telangiectasia. This

disparity is consistent with the idea that telangiectasia only develop under very specific conditions: likely that the somatic mutation must occur in a specific type of vascular bed, in an endothelial cell, and must be followed by local angiogenic stimulus.

2.4.6 Two-Hit Mechanism for *SMAD4* & JP-HHT

Our samples consisted of telangiectasia from individuals with HHT from a single HHT Centre of Excellence. This cohort had germline mutations in either *ENG* or *ACVRL1*, and somatic mutations were identified in telangiectasia from both genotypes. Mutations in *SMAD4* cause the combined syndrome HHT and Juvenile Polyposis (JP-HHT) , however individuals with *SMAD4* mutations only account for 2% of HHT cases.(?) Unfortunately, no individuals with JP-HHT were present in our cohort, but we believe it is likely that JP-HHT-related telangiectasia follow an identical genetic two-hit mechanism, resulting from somatic mutations in *SMAD4*.

2.4.7 Compound Heterozygosity

As HHT is caused by germline mutations in any one of 3 genes, an interesting question is whether the compound effect of a LOF mutation in two different HHT genes could drive pathogenesis (e.g. a germline mutation in *ENG* and a somatic mutation in *ACVRL1*). However, thus far all of the somatic mutations we identified in HHT-related telangiectasia occur in the same gene as that harboring the germline mutation. This human genetic data is consistent with the observation that mice with combined deficiency for one allele each of *Acvrl1* and *Eng* are fertile, viable and are not teeming with vascular malformations(?) (Srinivasan and Marchuk, unpublished), as might be expected if trans-heterozygosity of mutations in these two HHT-genes could initiate vascular malformation development.

2.4.8 Therapeutic Potential

A genetic two-hit mechanism for HHT pathogenesis has therapeutic implications. Certain efforts to develop therapies for HHT assume a model of haploinsufficiency of the relevant HHT gene. These strategies attempt to increase the amount of the affected gene product by increasing the level of transcript/protein arising from the wild-type allele.(?) We show here that some fraction of cells within the malformation do not possess a wild-type allele. Even if expression in surrounding heterozygous cells could be increased, the null cells would remain devoid of protein, suggesting that this avenue of therapy may be ineffective. By contrast, a more effective strategy may be gene replacement, reintroducing a fully wild-type allele into the mutated cells(?), as this would simultaneously provide an extra copy of the gene to both the heterozygous and null cells. This strategy is particularly attractive as it might inhibit new VM formation by adding back a second wild-type copy of the mutated gene.

2.5 Methods

Sample Collection

Individuals were enrolled in the study after giving informed consent (approved by either the St. Michael's Hospital IRB committee or the Duke University Health System IRB Committee). Diagnosis of HHT was based on identification of a pathogenic germline mutation, or by exhibiting at least three of the four symptoms as per the Curaçao criteria (Supp Table 1).(?) Telangiectasia were resected using a 3mm punch biopsy, after local anesthesia (1% xylocaine with epinephrine), with standard aseptic technique. Sample 6005-1 was immediately formalin fixed (10% formalin) and paraffin embedded (FFPE), and then shipped at room temperature. All other samples were immediately frozen at -80 Celsius, and then shipped on dry ice. Saliva samples were obtained using Oragene DNA saliva kits at the time of tissue collection.

Blood from individual 6003 was obtained during a subsequent visit, shipped at room temperature, and immediately used for RNA extraction.

DNA and RNA Extraction

DNA from telangiectasia samples was extracted using the DNeasy Blood and Tissue Kit (Qiagen). DNA from FFPE sample 6005-1 was extracted using QIAamp DNA FFPE Tissue Kit (Qiagen). Genomic DNA and RNA were extracted from peripheral blood leukocytes from individual 6003 and from a non-HHT control individual using the Gentra PureGene Blood Kit (Qiagen) and TRIzol Reagent (Invitrogen) extraction protocols, respectively, as per the manufacturers' directions.

Targeted Sequencing

To enable the detection of somatic mutations in telangiectasia we used a next generation sequencing strategy. Somatic mutations involved in the pathogenesis of other vascular malformation diseases such as cerebral cavernous malformations often have a low allele frequency due to somatic mosaicism in the malformation. Somatic mosaicism is also present in retinal AVMs from mouse models of HHT suggesting that low allele frequency may be a confounder when identifying somatic variants in telangiectasia. In addition, the telangiectasia samples collected for this study consist of bulk biopsied tissue which have not been enriched for any particular cell type. Considering these potential sources of normal (non-mutant) cell contamination, we sequenced telangiectasia to >1000x coverage and incorporated a unique molecular identifier to enable the detection of variants as low as 0.1% allele frequency. Eighteen fresh-frozen samples and one FFPE sample were sequenced using a custom Agilent SureSelect panel covering 16 genes implicated in various vascular malformation disorders: *ENG*, *ACVRL1*, *SMAD4*, *BRAF*, *CCM2*, *FLT1*, *FLT4*, *GNAQ*, *KDR*, *KRAS*, *KRIT1*, *MAP2K1*, *NRAS*, *PDCD10*, *PIK3CA*, and *PTEN*. Though *GDF2*

variants have been identified in an HHT-like phenotype, these cases are extremely rare. Moreover, mutations in GDF2 have not been identified in any individual at the Toronto HHT Centre for Excellence therefore *GDF2* was not included in the panel. To ensure the generation of high-quality sequencing libraries, the Agilent NGS FFPE QC Kit was used to determine the extent of DNA degradation in the FFPE sample. Samples with ΔCq values >2 were excluded from the study as per manufacturer recommendation. Sequencing libraries were generated using the Agilent SureSelect XT HS Kit. Samples were then pooled and sequenced on an iSeq 100 (Illumina) with paired-end 150bp reads. Across all samples, target regions were sequenced to a mean depth of 2803x with 78% of the target region at $>1000x$ and 96% of the target region at $>100x$.

Mutation Detection

Sequencing data was processed and analyzed using a custom pipeline based on the GATK best practices for somatic short variant discovery. Briefly, after analyzing the raw data with fastQC to ensure high quality data, the adapter sequences were trimmed from reads using bbduk, reads were aligned to the hg19 human reference genome using bowtie2, duplicates were removed based on UMI sequence using fgbio, variants were called using MuTect2 in tumor-only mode, and variants were annotated using snpSift. The resulting variant call file (VCF) was filtered through several steps to identify somatic mutations. To identify variants that may change the protein sequence or impact splicing, we selected for variants that occur within exons or within 10bp of an exon. From this set, we removed variants that are present in the population at $>0.01\%$ frequency by comparing to 3 databases (dbSNP, 1000 Genomes project, and the Exome Aggregation Consortium [ExAC]); as these variants are more common than the frequency of HHT (1–5 in 10000). (?) We removed any variants present in $<0.02\%$ of sequence reads, as this is the reported technical limit

of detection for the SureSelect XT technology. We also removed variants in regions with <100x coverage, variants with <5 supporting reads, variants that were strand specific, and variants where <50% of alternative bases had a quality score >30 (>Q30). Candidate somatic variants identified in the targeted sequencing data were then validated by sequencing amplicons generated during a second, independent round of PCR amplification. We designed primers for each sample to specifically amplify the position of the somatic mutation and 100–200bp of flanking sequence. When possible, the primers were designed such that they would capture the position of both the germline and somatic mutations within a single amplicon. Each primer was synthesized with the following Illumina flow cell adapter sequences such that the amplicons could be easily indexed and sequenced:

For: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-[primer]-3'

Rev: 5'-GTCTCGTGGCTCGGAGATGTGTATAAGAGACAG-[primer]-3'

Amplicons were prepared from telangiectasia DNA and constitutional DNA if available (Supp Table 1) using two rounds of PCR. The first round of PCR was 25 cycles and served to amplify the target region from genomic DNA. Amplicons from the first round were purified using AMPure XP beads (Beckman Coulter) and used for a second round of PCR using 8 cycles to attach a sample index using the Nextera XT Index Kit. Amplicons from the second round were purified, pooled, and sequenced on an iSeq 100 with 150bp paired-end reads to a depth of >10000x. The frequency of the somatic mutation in telangiectasia and constitutional DNA was determined using custom scripts, excluding bases <Q15. As these amplicons were sequenced in the same run, it is possible for a low level of index misassignment causing switching of reads between samples. This could cause some reads from telangiectasia to be assigned as constitutional reads and vice versa. To estimate the rate of index

misassignment, we examined pairs of samples that target different genomic locations and quantified the proportion of misassigned reads between these samples. Based on this we estimate that the rate of misassignment is 0.2–0.8%, relative to the sample of origin. For example, if one sample has a somatic mutation with a frequency of 1% and was sequenced to 100000x coverage, then 2–8 reads containing the somatic mutation would be misassigned to each of the other samples in the pool.

Establishing Phase

To establish the phase of the somatic and germline mutations we used either short-read sequencing with Illumina chemistry, or long-read sequencing with PacBio chemistry depending on the distance between the two mutations. For mutations <500bp apart, we generated amplicons during the validation process that cover the positions of both the somatic and germline mutations. These amplicons were sequenced on an iSeq 100 as described above. For mutations >500bp we designed primers such that the resulting amplicon would cover the position of both the germline and somatic mutations for each sample. Each primer was synthesized with the PacBio ‘universal tag’ as follows to enable indexing and sequencing:

For: 5’-/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC-[primer]-3’

Rev: 5’-/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG-[primer]-3’

We generated amplicons spanning the somatic and germline mutations using the LongAmp Taq DNA Polymerase Kit as per manufacturer instructions. These amplicons were purified with AMPure XP beads and used for a second round of PCR to attach the sample index. This process used no more than 30 cycles of PCR total. These amplicons were pooled and sequenced across one SMRT cell on a PacBio Sequel System. The sequence reads were aligned to the hg19 human genome using Minimap2

in ava-pb mode. The single molecule resolution of these technologies allowed us to determine how the mutant alleles are arranged; if the mutations are in trans then reads will have either the somatic mutant allele or the germline mutant allele, if the mutations are in cis then reads will have either no mutant alleles or both mutant alleles. In total we generated mutation-spanning reads for 7 telangiectasia, each with >100 reads which contained the somatic mutation. The p-values reported for phase status were calculated using a binomial distribution with the null hypothesis that a random mutation has an equal probability of cis or trans configuration with a nearby variant. The genomic distance between mutations varied greatly with the closest mutations in sample 6005-1 with 26 bases between mutations, and the most distant in 6001-10 with 18.7 kilobases between mutations. Of the 9 telangiectasia with identified somatic mutations, the distance between mutations in 6002-2, 6003-1, and 6005-1 was small enough to allow for mutation-spanning reads using illumina chemistry (Table. 2). Mutation-spanning reads were generated for 6001-3, 6001-7, 6001-8, and 6002-1 using PacBio chemistry. We were unable to generate amplicons spanning the mutations for 6001-1 and 6001-10. The sequence downstream of the somatic mutation in these telangiectasia contains several repetitive regions which, combined with the genomic distance and limited quantity of input DNA may have contributed to PCR failure. One notable confounder in this analysis is the generation of chimeric reads resulting from template switching during PCR. The generation of chimeric reads is known to interfere with amplicon-based haplotype phasing by switching a variant from one strand to another, potentially generating new haplotypes not present in the original sample.²³ In practice, chimeric reads randomize the arrangement of the somatic and germline mutations. The frequency of chimeric arrangements is highly dependent on the distance between mutations and the number of PCR cycles used to make the amplicons. To reduce the number of chimeric reads in our libraries we used no more than 30 cycles for amplification. A previous study reports a chimeric

arrangement frequency of 6.5% for 29 cycles of amplification for mutations 9kbp apart.(?) Chimeric reads may account for the very few discordant reads in some of our samples, as shown in Table 2. Nonetheless, these were so minor in comparison to the great majority of the reads that phase could be unequivocally determined.

in vitro Splicing

A 3.8kb fragment of *ACVRL1* genomic DNA spanning from 431 bases upstream of exon 3 to 215 bases downstream from exon 8 was amplified and this entire insert was ligated into the MCS of pSPL3, a splicing vector.(?) Clones were sequenced to ensure that no PCR-induced errors were present in the exons and adjacent intronic regions of the insert. The specific mutation, c.625+4A>T, was introduced using site directed mutagenesis and again, clones were sequenced to verify that the only sequence difference was at the intended site. Plasmid DNA from empty vector, wild type (control) vector and mutation-containing vector were transfected into HEK293T cells using Lipofectamine 3000 (ThermoFisher Scientific), incubated for 24 hours, and then the RNA was extracted using TRIzol and Direct-zol RNA miniprep kit (Zymo Research).

Reverse-Transcription PCR

RNA extracted from peripheral blood leukocytes and from transfected cells was used as template for cDNA synthesis using the Maxima H Minus First Strand cDNA kit (ThermoFisher Scientific). An RT primer in *ACVRL1* exon 8 was used in the RNA from blood while a vector-specific RT primer was used for the RNA from transfected cells to ensure that only RNA from the transfected vectors was being used as template for cDNA synthesis. cDNA from peripheral blood leukocytes was PCR amplified using primers in exons 3 and 6 while cDNA from the transfected cells was PCR amplified using primers in exons 3 and 8. PCR reactions were run on 1%

agarose gels, the bands excised and Sanger sequenced.

2.6 Acknowledgements

This study has been published in *AJHG* (?) with the following authors: Daniel A. Snellings, Carol J. Gallione, Dewi S. Clark, Nicholas T. Vozoris, Marie E. Faughnan, and Douglas A. Marchuk. We thank all participants and their families who donated samples for this study. We thank Dr. Nicolas Devos and the Duke Sequencing and Genomic Technologies Core for assistance designing experiments and sequencing. This study was supported by U.S. Department of Defense grant (W81XWH-17-1-0429) and a Fondation Leducq Transatlantic Network of Excellence Grant in Neurovascular Disease (17 CVD 03). MEF also received financial support from the Nelson Arthur Hyland Foundation.

3

Mutant *GNAQ* Alleles Produce Distinct Disease Phenotypes

3.1 Premise

3.2 Results

3.2.1 *Port Wine Stain with Uncommon Mutation GNAQ p.Q209R*

3.2.2 *Structural Analysis of Common GNAQ Alleles*

3.2.3 *Functional Analysis of Common GNAQ Alleles (???Not sure it is appropriate to add this???)*

3.2.4 *Transcriptional Analysis of Common GNAQ Alleles*

3.3 Discussion

3.3.1 *Distinct Functions of GNAQ Alleles May Underlie Disease Specificity*

3.3.2 *Relationship Between GNAQ Activity and Mutation Timing*

3.3.3 *Importance of Mutated Cell Type in Determining Disease Fate*

3.4 Methods

4

MAP3K3 Mutations Seed Cerebral Cavernous Malformations

4.1 Premise

4.2 Results

4.2.1 *MAP3K3 Somatic Mutations Only Occur in Sporadic CCM*

4.2.2 *MAP3K3 and CCM Gene Mutations are Mutually Exclusive*

4.2.3 *Mutations in KLF4 Do Not Contribute to CCM*

4.2.4 *(Whole-Exome Results) (??? Merge with above ???)*

4.3 Discussion

4.3.1 *CCM Loss of Function and MAP3K3 Gain of Function are Functionally Equivalent*

4.3.2 *Differing Constraints for Constitutional and Somatic Inheritance*

4.4 Methods

5

PIK3CA Mutations Fuel Cerebral Cavernous
Malformation Growth

5.1 Premise

5.2 Results

5.2.1 *PIK3CA Mutations Occur in Familial and Sporadic CCMs*

5.2.2 *CCMs Harbor Multiple Somatic Mutations in Different Genes*

5.2.3 *PIK3CA and CCM/MAP3K3 Mutations in the Same Cell*

5.2.4 *Developmental Venous Anomalies Predispose to Malformation*

5.3 Discussion

5.3.1 *Three-Hit Model of CCM Pathogenesis*

5.3.2 *Similarities to the Genetic Mechanism of Cancer*

5.3.3 *Role of Clonal Expansion in Mutagenesis*

5.3.4 *Therapeutic Implications*

5.3.5 *Distinct Properties of PIK3CA vs. CCM/MAP3K3 Mutations*

5.3.6 *DVA Predispose to CCM and Other PI3K-Related Diseases*

5.4 Methods

CCM Collection

Brain AVM Collection

DNA Extraction

Droplet Digital PCR

SNaPshot

Sequencing

Sequence Analysis

Single-Nucleus DNA Sequencing

Statistics

6

Developmental Venous Anomalies Predispose to Malformation

6.1 Premise

6.2 Results

6.3 Discussion

6.4 Methods

Conclusions & Musings

7.1 HHT Pathogenesis

7.2 CCM Pathogenesis

7.2.1 *Regrowth after Surgical Resection*

7.2.2 *CCM & Meningioma*

Although the vascular lesions are the primary sequelae of familial CCM, many groups have noted an increased prevalence of meningioma in individuals with familial CCM—especially those with a mutation in *PDCD10* (Labauge et al., 2009; Riant et al., 2013; Garaci et al., 2015). In addition, we have previously been contacted by an individual whose child had a sporadic CCM that regrew into a meningioma. Unfortunately we were unable to acquire a tissue sample for genetic analysis, however this case along with the strong link between familial CCM and meningioma has fueled my interest in understanding the link between CCM and meningioma.

Kruppel-like factor 4 (KLF4)

KLF4 is a transcription factor with key roles in the pathogenesis of both CCM and meningioma. In CCM, *KLF4* (along with *KLF2*) is a key player in CCM

signaling that is upregulated by loss of the CCM complex or gain of function in *MAP3K3* (Cuttano et al., 2016; Zhou et al., 2016). Indeed, overexpression of *KLF4* in mouse models leads to an aggressive CCM-like phenotype (Ren et al., 2021). In meningioma, mutations in *KLF4* are common and often co-occur with mutations in *TRAF7* (Reuss et al., 2013). In addition, *KLF4* mutated meningioma have been shown to be responsive to treatment with temsirolimus (von Spreckelsen et al., 2020)—a derivative of rapamycin (sirolimus) which has been shown to be effective in treating CCM in mice (Ren et al., 2021). The importance of *KLF4* in both CCM and meningioma suggests that dysregulation of *KLF4* may underlie the development of meningioma in individuals with familial CCM.

The majority of *KLF4* mutations found in the catalog of somatic mutations in cancer (COSMIC) result in p.K409Q (Figure 7.1A). This narrow spectrum of mutations suggests that the p.K409Q variant results in *KLF4* gain of function. Notably, the p.K409Q mutation is highly specific to meningioma, occurring in very few other cancer types (Figure 7.1B). The *KLF4* protein contains 3 DNA-binding zinc-finger domains that contribute to its activity as a transcription factor. The p.K409Q mutation occurs in the first of the three zinc-finger domains (Figure 7.1C). A structural study of *KLF4* determined that the second two zinc-finger domains consistently bind DNA, however the first zinc-finger only sometimes participates in DNA binding (Schuetz et al., 2011). They determined that the second and third domains are required for site specificity, but the third "inhibits cryptic self-renewal and block of differentiation activity". This immediately suggests a mechanism by which the p.K409Q mutation may disrupt the inhibitory capacity of *KLF4* while retaining its functions that are independent of the first zinc-finger domain.

I thought that since overexpression of *KLF4* is sufficient to cause CCM in mice, the p.K409Q mutation may cause some sporadic CCMs in humans. To test this, I developed a ddPCR assay to test for this mutation in 30 mutation-negative sporadic

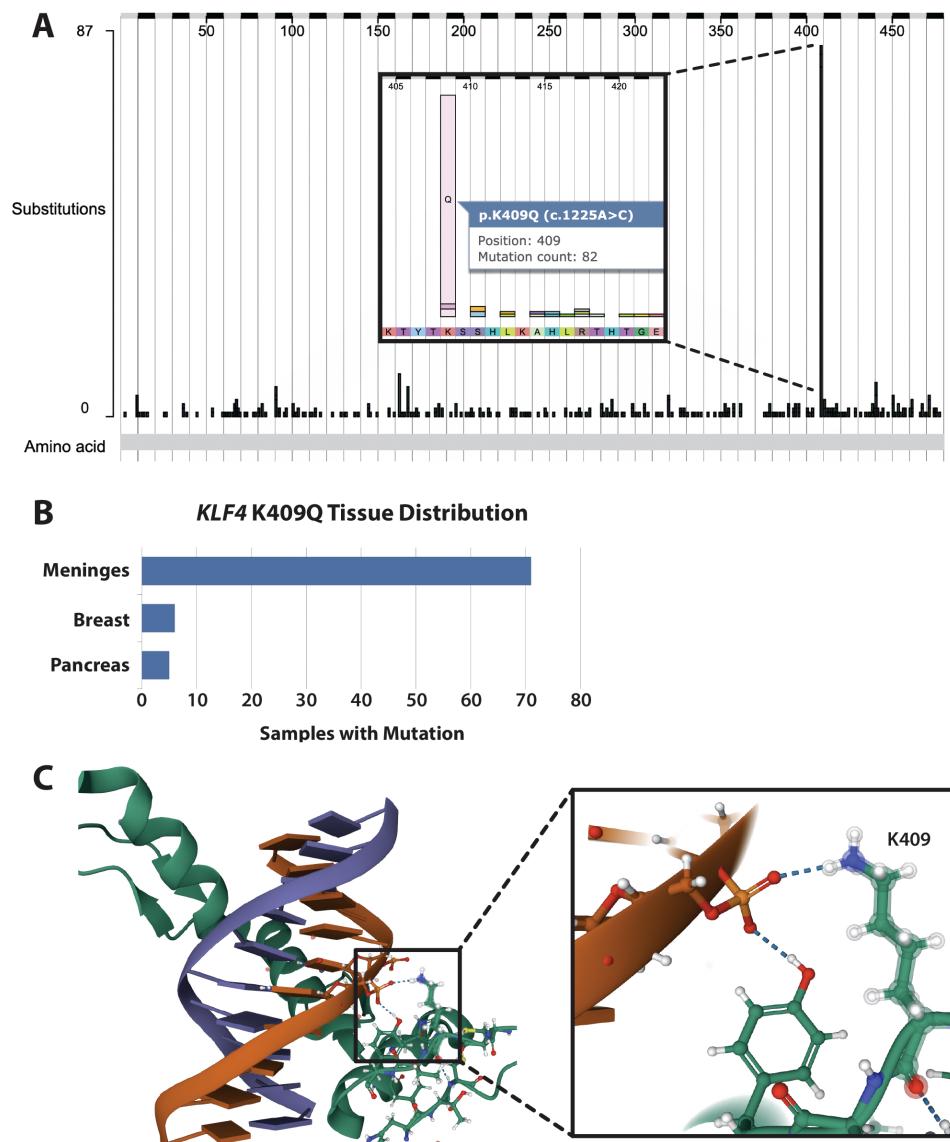


FIGURE 7.1: *KLF4* Mutations in COSMIC.

A. Distribution of somatic mutations in *KLF4* that are present in the catalog of somatic mutations in cancer (COSMIC). The location of the most frequent mutation (p.K409Q) is expanded in the inset. B. Distribution of *KLF4* p.K409Q in tissue types present in COSMIC. C. Structure of *KLF4* zinc-finger domains bound to DNA (pdb: 2WBU). Border box denotes the expanded region to the right showing the interaction between K409 and the DNA backbone.

CCM, however I did not find evidence for this mutation in any of these samples. This may be a reflection of the highly specific functions of the first zinc-finger domain in *KLF4* as discussed above. This function may be critical for meningioma development, but less important for CCM development. However, the inverse is not necessarily true—i.e. while the p.K409Q mutation may not cause CCMs, overexpression of *KLF4* may cause meningioma. This is currently unknown, though if true, may account for the presence of meningioma in familial CCM. One straightforward way to test this hypothesis would be to sequence tissue from meningioma in individuals with familial CCM. I suspect that these meningioma harbor a somatic loss of function mutation in *KRIT1/CCM2/PDCD10* leading to biallelic loss of a CCM gene resulting in upregulation of *KLF4*. These meningioma may also harbor a second mutation in *TRAF7* which have often been found in non-CCM meningioma. Unfortunately, tissue samples of meningioma from individuals with familial CCM have been difficult to acquire, thus I have been unable to test this hypothesis. However, I expect that future studies will find that *KLF4* is a critical link between CCM and meningioma.

7.3 Developmental Venous Anomalies as a Primer for Disease

7.3.1 *Association with Sporadic CCM*

7.3.2 *Association with Other Diseases*

7.3.3 *Cowden Syndrome*

7.3.4 *Implications*

7.4 Other Vascular Malformations

7.4.1 *Classification of Vascular Malformations and Vascular Tumors*

7.4.2 *Sturge-Weber Syndrome and Somatic Mutations in GNAQ*

Happle's hypothesis

GNAQ in uveal melanoma and circumscribed choroidal hemangioma

Link between SWS, UM, and CCH?

7.4.3 *The Curious Case of Infantile Hemangioma*

Infantile hemangioma (IH) are one of the most common vascular malformations.

They occur in children and are typically present at birth as a red spot flush with the surrounding skin. Soon after birth the hemangioma rapidly grows and becomes raised from the skin. They are generally benign and are typically left alone unless they cover the child's mouth, nose, or eyes. What makes IH so interesting compared to other types of vascular malformations is that they almost always completely regress within the first few years of the child's life. While other types of vascular malformations may spontaneously regress (telangiectasia and AVMs), none do so with the consistency of IH. This phenomenon has been of great interest, not for the purpose of developing therapeutics for IH (propranolol is an extremely effective treatment for IH) but for uncovering the mechanism of regression in the hopes that what we learn can be applied to regress other, more nefarious, vascular malformations.

GLUT1 in IH endothelium

Perhaps one of the most provocative discoveries into the mechanism of IH pathogenesis is the fact that endothelial cells from IHs highly express GLUT1 (North et al., 2000, 2001). GLUT1 is a glucose transporter that has remarkable specificity for the placental endothelium. This finding suggested that the IH may be comprised of cells that dislodged from the maternal placenta, then became hyper-proliferative in a post-fetal environment. If this hypothesis is correct, one would expect to find that the IH is a genetically chimeric growth between fetal and maternal cells. This hypothesis was put to the test using fluorescence *in situ* hybridization to assay the presence of XX cells in IH from a male infant with confirmation by sequencing microsatellites and SNPs that were divergent between mother and child. This analysis found no evidence for maternal-fetal chimerism in IH (Pittman et al., 2006). Despite this counter-evidence, the presence of GLUT1 in IH is strongly indicative of some link with the placenta though unfortunately this link currently remains elusive. The current literature has very clearly shown the presence of GLUT1 in IH endothelial cells, however the extent of the placental transcriptional program is unclear. Epigenomic profiling of paired IH and maternal placental samples may give valuable insights into the mechanism of IH.

Efforts to find somatic mutations in IH

As it is quickly becoming clear that the vast majority of vascular malformations are the result of somatic mutations—many occurring in known oncogenes—I thought that somatic mutations may also underlie IH. To test this, I sequenced 61 IH lesions on an ‘oncopanel’ covering many genes that are highly mutated in cancers as well as several genes previously implicated in vascular malformations (*KRIT1*, *CCM2*, *PDCD10*, *ACVRL1*, *ENG*, *SMAD4*, etc.) Unfortunately after filtering putative variants, no there were no variants with likely functional significance and that occurred in

more than a single sample. I am aware of at least 1 other group that has attempted to identify somatic mutations in IH via whole-exome sequencing, however to date there are no known somatic mutations in IH. One important aspect of these studies that must be noted is that they are invariably focused on coding regions of the genome. Non-coding variants are more than capable of causing disease however discovery-focused sequencing studies often ignore non-coding regions both because of the cost of sequencing the entire genome to a depth sufficient to detect somatic mutations, and the challenges associated with functional analysis of non-coding variants. Further studies may find that somatic mutations do cause IH, but they occur in a region of the genome that is missed by the majority of sequencing studies.

7.5 The Molecular Basis of Genetic Dominance

7.5.1 *Phenotypic Dominance ≠ Genetic Dominance*

7.5.2 *Knudsons Fingerprint*

7.5.3 *The Diverse Functional Effects of Genetic Mutations*

7.6 The Intersection of Somatic Mutagenesis and Evolution

7.6.1 *The Creation of New Alleles*

7.6.2 *The Relationship Between Mutability and Fitness Landscape*

7.6.3 *Recurring Mutations & Convergent Evolution*

7.6.4 *Clonal Evolution of Somatic Mutants*

7.6.5 *Cancer*

7.7 Somatic Mutations

7.7.1 *The Role of Somatic Mutations in Aging*

7.7.2 *Constitutional Intolerance & Somatic Permissiveness*

7.7.3 *Somatic Reversion of Pathogenic Mutations*

7.7.4 *What is the Consequence of RNA Mutations?*

7.8 Innovation in the Sequencing Era

7.8.1 *Detection of Somatic Mutations*

7.8.2 *Single Cell Sequencing*

7.8.3 *Utility of Rare Disease Research in Mechanistic Discovery*

7.8.4 *Data Democratization & Individual Privacy*

7.8.5 *Growing Importance of Informatics in Biology*

Appendix A

Probability of Multiple Somatic Mutations

Bibliography

- Cuttano, R., Rudini, N., Bravi, L., Corada, M., Giampietro, C., Papa, E., Morini, M. F., Maddaluno, L., Baeyens, N., Adams, R. H., Jain, M. K., Owens, G. K., Schwartz, M., Lampugnani, M. G., & Dejana, E. (2016). KLF4 is a key determinant in the development and progression of cerebral cavernous malformations. *EMBO Mol Med*, *8*(1), 6–24.
URL <https://www.ncbi.nlm.nih.gov/pubmed/26612856>
- Garaci, F., Marsili, L., Riant, F., Marziali, S., Cecillon, M., Pasquarelli, R., Sangiuolo, F., Floris, R., Novelli, G., Tournier-Lasserve, E., & Brancati, F. (2015). Cerebral cavernous malformations associated to meningioma: High penetrance in a novel family mutated in the PDCD10 gene. *Neuroradiol J*, *28*(3), 289–93.
URL <https://www.ncbi.nlm.nih.gov/pubmed/26246098>
- Labauge, P., Fontaine, B., Neau, J. P., Bergametti, F., Riant, F., Blecon, A., Marchelli, F., Arnoult, M., Lannuzel, A., Clanet, M., Olschwang, S., Denier, C., & Tournier-Lasserve, E. (2009). Multiple dural lesions mimicking meningiomas in patients with CCM3/PDCD10 mutations. *Neurology*, *72*(23), 2044–6.
URL <https://www.ncbi.nlm.nih.gov/pubmed/19506228>
- North, P. E., Waner, M., Mizeracki, A., & Mihm, J., M. C. (2000). GLUT1: a newly discovered immunohistochemical marker for juvenile hemangiomas. *Hum Pathol*, *31*(1), 11–22.
URL <https://www.ncbi.nlm.nih.gov/pubmed/10665907>
- North, P. E., Waner, M., Mizeracki, A., Mrak, R. E., Nicholas, R., Kincannon, J., Suen, J. Y., & Mihm, J., M. C. (2001). A unique microvascular phenotype shared by juvenile hemangiomas and human placenta. *Arch Dermatol*, *137*(5), 559–70.
URL <https://www.ncbi.nlm.nih.gov/pubmed/11346333>
- Pittman, K. M., Losken, H. W., Kleinman, M. E., Marcus, J. R., Blei, F., Gurtner, G. C., & Marchuk, D. A. (2006). No evidence for maternal-fetal microchimerism in infantile hemangioma: a molecular genetic investigation. *J Invest Dermatol*, *126*(11), 2533–8.
URL <https://www.ncbi.nlm.nih.gov/pubmed/16902414>

- Ren, A. A., Snellings, D. A., Su, Y. S., Hong, C. C., Castro, M., Tang, A. T., Detter, M. R., Hobson, N., Girard, R., Romanos, S., Lightle, R., Moore, T., Shenkar, R., Benavides, C., Beaman, M. M., Mueller-Fielitz, H., Chen, M., Mericko, P., Yang, J., Sung, D. C., Lawton, M. T., Ruppert, M., Schwaninger, M., Korbelin, J., Potente, M., Awad, I. A., Marchuk, D. A., & Kahn, M. L. (2021). *PIK3CA* and CCM mutations fuel cavernomas through a cancer-like mechanism. *Nature*. URL <https://www.ncbi.nlm.nih.gov/pubmed/33910229>
- Reuss, D. E., Piro, R. M., Jones, D. T., Simon, M., Ketter, R., Kool, M., Becker, A., Sahm, F., Pusch, S., Meyer, J., Hagenlocher, C., Schweizer, L., Capper, D., Kickingereder, P., Mucha, J., Koelsche, C., Jager, N., Santarius, T., Tarpey, P. S., Stephens, P. J., Andrew Futreal, P., Wellenreuther, R., Kraus, J., Lenartz, D., Herold-Mende, C., Hartmann, C., Mawrin, C., Giese, N., Eils, R., Collins, V. P., Konig, R., Wiestler, O. D., Pfister, S. M., & von Deimling, A. (2013). Secretory meningiomas are defined by combined *KLF4* K409Q and *TRAF7* mutations. *Acta Neuropathol*, 125(3), 351–8. URL <https://www.ncbi.nlm.nih.gov/pubmed/23404370>
- Riant, F., Bergametti, F., Fournier, H. D., Chapon, F., Michalak-Provost, S., Cecillon, M., Lejeune, P., Hosseini, H., Choe, C., Orth, M., Bernreuther, C., Boulday, G., Denier, C., Labauge, P., & Tournier-Lasserve, E. (2013). CCM3 Mutations Are Associated with Early-Onset Cerebral Hemorrhage and Multiple Meningiomas. *Mol Syndromol*, 4(4), 165–72. URL <https://www.ncbi.nlm.nih.gov/pubmed/23801932>
- Schuetz, A., Nana, D., Rose, C., Zocher, G., Milanovic, M., Koenigsmann, J., Blasig, R., Heinemann, U., & Carstanjen, D. (2011). The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. *Cell Mol Life Sci*, 68(18), 3121–31. URL <https://www.ncbi.nlm.nih.gov/pubmed/21290164>
- von Spreckelsen, N., Waldt, N., Poetschke, R., Kesseler, C., Dohmen, H., Jiao, H. K., Nemeth, A., Schob, S., Scherlach, C., Sandalcioglu, I. E., Deckert, M., Angenstein, F., Krischek, B., Stavrinos, P., Timmer, M., Remke, M., Kirches, E., Goldbrunner, R., Chiocca, E. A., Huettelmaier, S., Acker, T., & Mawrin, C. (2020). KLF4(K409Q)-mutated meningiomas show enhanced hypoxia signaling and respond to mTORC1 inhibitor treatment. *Acta Neuropathol Commun*, 8(1), 41. URL <https://www.ncbi.nlm.nih.gov/pubmed/32245394>
- Zhou, Z., Tang, A. T., Wong, W. Y., Bamezai, S., Goddard, L. M., Shenkar, R., Zhou, S., Yang, J., Wright, A. C., Foley, M., Arthur, J. S., Whitehead, K. J., Awad, I. A., Li, D. Y., Zheng, X., & Kahn, M. L. (2016). Cerebral cavernous malformations arise from endothelial gain of MEKK3-KLF2/4 signalling. *Nature*,

532(7597), 122–6.

URL <https://www.ncbi.nlm.nih.gov/pubmed/27027284>

Biography

Your biography is limited to one page and must contain

1. Full name
2. Date and place of birth
3. Every degree you've earned, including this one, and where you earned it from.

Mostly, that information is to narrow down which John Smith wrote that dissertation on the mating habits of sea cucumbers. Sexy!

You may also include

1. Any awards you've won related to your discipline since your undergraduate degree.
2. Any fellowships you've held
3. Anything you've published (papers, books, book chapters). Don't be afraid to cite it here, so that the full bibliographic record of your article appears in the bibliography!
4. Where your next job will be, if you know