Allele-specific analysis of genomic data reveals novel cancer and lncRNA biology

Michael B. Heskett

Dissertation

Presented to the Department of Molecular and Medical Genetics

Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

September 2021

School of Medicine

Oregon Health & Science University

DOCTORAL DISSERTATION APPROVED BY DISSERTATION ADVISORY COMMITTEE:

Dr. Paul Spellman

Dr. Matt Thayer

Dr. Andrew Adey

Dr. Laura Heiser

Dr. Abhi Nellore

NOVEMBER 27, 2021.

TABLE OF CONTENTS

THIS PAGE LEFT BLANK

Acknowledgements

Thank you to everyone that helped to make my research and career development possible and offered guidance along the way. Thanks to: Dr. Spellman for unwavering support of my career development and for many stimulating conversations of genetics and the scientific method; Dr. Thayer for sharing with me his passion and excitement for RNA and chromosome biology; Drs. Adey, Heiser, and Nellore for serving on my dissertation committee and offering very helpful advice throughout all the graduate school milestones; Drs. Dai and McCullough for organizing the Genetics graduate program and helping navigate all the twists and turns of graduate school; Dr. Moore for facilitating the great environment in which we held journal club for many years; Dr. Maslen for so much guidance, good advice, and facilitation of the PMCB courses and the PERT training program which allowed me a strong start to graduate school; all the folks in the MMG administrative office who helped keep the wheels turning; Melissa Brewer for expertly walking me through my Fellowship applications; The Knight Cancer Institute for hosting me; NIH/NCI for assessment and feedback of my research proposal and funding my work; Dr. Gilbert at FSU for a great collaboration; Drs. Solomon and Cho for another great collaboration; all of my fellow colleagues and trainees at OHSU PMCB who took courses with me, talked about science any time, and contributed to a positive and friendly environment at OHSU; and finally my family Marcia, Curtis, and Megan for unconditional support.

Dissertation Introduction

The dissertation work presented below covers two main areas of interest: genetic mutations in ovarian teratoma, and the discovery of a new member of the ASAR family of long non-coding RNAs that is essential for normal DNA replication timing and chromosome stability. In ovarian immature teratoma, a malignancy of the female reproductive tract that affects young women, we have utilized exome sequencing of a cohort of 10 cancer patients to detect somatic mutations and learn the genetic mechanism of tumor origination as well as the genetic relationship between multiple tumor components within individual cancer patients. We present the largest modern genomic study of the immature teratoma histological subtype of ovarian germ cell cancer to date, and demonstrate the very low somatic mutation rate, diploid genomes, and widespread copy-neutral loss of heterozygosity. The research presented here will lay the groundwork for future researchers to develop the potential of genetic biomarkers of clinical cancer behavior, as well as inform the field of mechanisms of meiotic abnormalities during germ cell development. In chapter II, the discovery of a novel long non-coding RNA that is required for normal DNA replication timing and stability of chromosome 6 is reported. DNA replication timing is an epigenetic phenomenon that describes the temporal order of replication along chromosomes, and is tightly linked with expression level, chromatin structure and subnuclear localization, and mutation rate, however insight into the mechanism of replication timing regulation is lacking in the literature. With the ASAR family of long non-coding RNAs we demonstrate that ASAR6-141, a non-canonical very long non-coding RNA, is expressed from chromosome 6 in a mono-allelic manner, associates with the parent chromosome homolog i*n cis*, and that disruption of the ASAR6-141 locus results in aberrantly delayed replication timing. The discovery of ASAR6-141

furthers the knowledge of basic chromosome functionality, as well as describes a new mechanism by which ageing cells could acquire chromosome instability, a hallmark feature of cancer. Future work in an effort to generalize the findings of ASAR6-141 to the whole genome is described in the conclusion section. The main area of overlap between the two research endeavors described above is the consideration of allele-specific genetic behavior, appropriate allele-specific analytic methods as critical challenges, and the development of biological knowledge on how the disruption of differential allelic processes can lead to outcomes such as genome instability and neoplastic growth.

Chapter 1: Allele-specific Analysis of Ovarian Teratoma Reveals Widespread Copy-Neutral Loss of Heterozygosity

1.1 Abstract

Ovarian teratoma, comprising cystic mature teratomas and malignant immature teratomas, is a common cancer of the female reproductive tract that tends to occur in young women, forms a relatively large tumor mass, and displays differentiated or primitive tissue from 2 or 3 germ cell layers. Ovarian teratomas contain highly heterogeneous cell morphology, may disseminate to distant tissues, and occur bilaterally in 10-15% of cases. Early stage immature teratoma and mature teratoma are highly susceptible to platinum-based chemotherapy and can be surgically removed with high success rates. Believed to be derived from germ cells, the genetics of ovarian teratomas, especially the malignant immature form, has not been well studied and the driving forces of teratoma origination and development is currently unknown. Here we utilized knowledge from testicular germ cell tumors to test hypotheses of the origin and development of ovarian teratomas including the presence of KIT and KRAS oncogene mutations and ubiquity of the isochrome 12p variant. By exome sequencing, we have found a unique tumor etiology that is defined by the absence of significant point mutations, but widespread copy-neutral loss of heterozygosity affecting thousands of genes in each tumor genome. We also utilize loss of heterozygosity patterns to infer the clonal relationship between morphologically contrasting regions of immature teratomas, and uncover strong evidence that failure of meiosis I, meiosis II, or whole genome duplication is the defining genetic event of ovarian teratoma.

1.2 Introduction

Owing to their striking and unusual morphology, cases of teratoma are easily recognized and have been found in medical texts as early as 1658, when German physician Johann Scultetus illustrated and described a tumor mass containing hair [1]. Teratoma, from Greek téras:*monster* was coined in 1863 by German pathologist Rudolf Virchow likely in reference to the common appearance of fully formed tissues including hair, teeth, and skin [2]. Teratoma belongs to the large and diverse group of germ cell tumors (GCTs) which contain malignant and benign neoplasm subgroups thought to be derived from the precursors of the gametes in both sexes, known as the germ cells.

Overview of germ cell tumors

Germ cell tumors of the female reproductive tract are divided into seven groups by the World Health Organization (WHO) based on histopathological features: dysgerminoma, yolk sac tumor, embryonal carcinoma, non-gestational choriocarcinoma, mature teratoma, immature teratoma, and mixed germ cell tumor [3](Figure 1A). Two additional general comparisons are made, dysgerminoma vs non-dysgerminoma, and malignant vs benign. Dysgerminoma is defined as a germ cell tumor being composed of cells showing no specific pattern of differentiation, and is the most common type of malignant ovarian germ cell tumor (Figure 1B). Incidence of teratoma peaks in young women aged 15-19 and slowly decreases with increased age, as opposed to other tumors of the female reproductive tract in which incidence increases over time (Figure 1C). Long term survival of dysgerminoma and teratoma is greater than 80%(Figure 1D). Non-dysgerminoma tumors are defined by a variety of distinctive patterns of differentiation, macroscopy,

histopathology, and epidemiology; however, the molecular characteristics and genetic profiles of each tumor subtype are not well studied. Germ cell tumors may be malignant or benign and the most frequently occurring of ovarian germ cell tumors is the mature teratoma, making up ~20% of all ovarian neoplasms and commonly referred to as a dermoid cyst, composed of mature tissues derived from two or more of the germ cell layers [3].

Homologous tumors occur in males, known as the testicular germ cell tumors, and are the most common malignancy in young men of European descent[4]. Analogous to the dysgerminoma vs non-dysgerminoma comparison, testicular germ cell tumors are classified into seminoma and nonseminoma groups, respectively, and also exhibit histopathological subtypes including yolk sac tumors, teratoma, embryonal carcinoma, and choriocarcinoma [5]. Both seminoma and nonseminoma have relatively high survival rates compared to other solid tumors owing to successful surgery and high sensitivity to modern chemotherapeutic drugs.

Germ cell tumors are unique from all other tumors of the body in that their cell type of origin is the germ cell, one that is specified early in embryonic development, undergoes a long and coordinated migration across embryonic tissues, undergoes the process of meiosis involving chromosome recombination, resides in an organ specially adapted to foster the development of these cells, and has totipotent capability to form a new individual (Figure 2A,2B,2C). Thus, to understand germ cell tumors we must first look at the development of the germ cells and their unique journey from primordial germ cell to gamete in both sexes.

Precursors of gametes are known as germ cells

At week 3 in human embryonic development, the primitive streak is formed and gives rise to the primordial germ cells which will eventually become gametes with the potential to form the next generation of individual. Because of ethical limitations, most knowledge of mammalian embryonic

development has been generated in the mouse. Primordial germ cells are distinguished from the embryo very early in development and follow a tightly orchestrated migration shortly after their specification. In mice, primordial germ cells establish cellular polarity and migrate from the primitive streak to the endoderm layer at embryonic day 7.5 (E7.5), along the endoderm at E8, followed by bilateral migration into the dorsal mesentery and lastly at E10.5 primordial germ cells reach the genital ridges and continue bilateral migration to reside in the embryonic gonad [6]. Migration into the genital ridge is dependent on SDF-1 expression in the genital ridge and CXCR4 expression within the migrating primordial germ cells [6]. After migration to the gonads, the population of primordial germ cells increases through mitotic division to several million in humans, and tens of thousands in mice. The KIT receptor tyrosine kinase and its ligand KITLG also widely appreciated for their roles in primordial germ cell proliferation, migration, and survival, where KIT is required for efficient migration through the hindgut[6]. In females, primordial germ cells undergo meiosis in the 12th week of human development and the cells are then referred to as primary oocytes. Mitotic division leads to the production of ~7million germ cells, however this population is greatly reduced before the first meiotic division of cells that will become oocytes. Primary oocytes are arrested in prophase I of meiosis and this arrest will last until puberty where each month until menopause the first meiosis is completed followed by meiosis II and ovulation[7]. Males are continuously fertile after puberty and undergo mitotic expansion of spermatogonia during adulthood. In males after pubertal age production of gametes occurs in distinct phases, that lasts in total about 64 days. Spermatogonia are undifferentiated germ cells that reside in the seminiferous tubules and undergo mitosis. One daughter cell will replenish the population of germ cells known as spermatogonia type A1, and the other cell known as spermatogonia type A2 will continue to divide for several generations by mitosis and migrate

towards the lumen in a process known as clonal expansion [7]. After several mitotic expansions of spermatogonia, the primary spermatocytes are formed which are the cells that will undergo meiosis to produce haploid gametes. During spermatogonal mitotic divisions, cytokinesis does not occur, forming large bridged structure known as the syncytium. Cells within the syncytium undergo meiosis I and II to produce spermatids with haploid nuclei, but remain connected and are functionally diploid as sharing of gene products can occur. The final stage of differentiation called spermatogenesis occurs producing ~100 million gametes per day in each human testicle [7].

Epidemiology of ovarian germ cell tumors

Over the last 30 years, malignant ovarian tumor incidence has slightly decreased to 0.338 cases per 100,000 women years [8]. In contrast, testicular malignant germ cell tumor incidence is increasing and is estimated at 7-8 per 100,000. 5-year survival as measured by the Surveillance, Epidemiology, and End Results (SEER) program between 1973-2002 is greater than 80% in dysgerminoma for women age 49 and younger [8]. The study of 1,262 patients found approximately one third of diagnoses were approximately evenly split between dysgerminoma, immature teratoma, and non-dysgerminoma subtypes. Data from the Thames Cancer Registry collected from 1960-1999 from patients in South East England found the incidence of male germ cell tumors was increased from 2.0 to 4.4 cases per 100,000 and there was a subtle increase in female incidence from ~0.15 to 0.25 per 100,000[9]. Both male and female germ cell tumors share the unusual feature that incidence peaks early in life. In men, non-seminoma is common early in life and seminoma later in life, with the overall peak incidence occurring around age 30. In females, highest incidence is in the late teens and dysgerminoma is common early in life whereas teratoma is most commonly seen later in life [9]. Although seminoma and dysgerminoma are homologous, their incidence early in females but late in males implies that sex specific differences are not trivial.

Germ cell tumors are highly sensitive to platinum based therapy

Long term survival of germ cell tumors is among the highest of any cancer type owing to success of surgery and remarkable sensitivity to platinum based chemotherapy and radiation. The cure rate of early stage malignant ovarian germ cell tumors is nearly 100%, and about 75% for advanced stage disease[10]. Cisplatin is able to crosslink with purine bases on DNA causing 1,2-instrastrand d(CpG) and 1,2-intrastrand d(ApG) adducts. Cancer cells with insufficient DNA repair may trigger apoptosis after exposure to cisplatin. Additionally, cisplatin may induce reactive oxygen species that trigger cell death independent of DNA damage pathways[11]. Cisplatin is one of the most widely used chemotherapies, as such many plausible mechanisms of actions are known, however, genetic and epigenetic factors that contribute to the unique sensitivity of germ cell tumors to cisplatin are unknown, representing a stark knowledge gap. Standard chemotherapy regimen for malignant ovarian germ cell tumors is a combination of bleomycin, etoposide, and cisplatin. Surgery alone is an attractive alternative for early stage disease and confers a relatively low additional risk while avoiding side effects of chemotherapeutic exposure. A Norwegian study of 351 patients diagnosed with malignant ovarian germ cell tumors found that 0/31 patients with 10-year survival who were treated with surgery alone eventually developed a second cancer, whereas 23/139 patients who survived 10 years but were originally treated with cytotoxic treatment and radiation eventually developed a second cancer [12]. Although many variables may confound this comparison, second cancer risk in young patients undergoing radiotherapy has been widely appreciated and underlines the need for biomarkers of tumor aggressiveness which would allow avoidance of cytotoxic and radiological therapy in some cases avoiding the increased risk of second cancer.

Predisposition to germ cell tumors is suspected in people with disorders of sex development

Disorders of sex development (DSD) refers to a group of conditions involving incomplete or disordered gonadal development. Three groups of DSD individuals show increased risk of germ cell tumors: 1) 46,XY with disorders of male development due to mutations in critical genes such as SRY or AR, 2) alterations in sex chromosomes such as 47,XXY and 45,X, and 3) 46,XX with female gonadal disorders due to various aberrations including presence of SRY or mutations in RSPO1 [10, 13]. Evaluations of several DSD patient series has led to reports of increased risk of germ cell tumors with a range of zero to 60%. Families with ovarian germ cell tumors have been reported, but are rare and only a few examples exist in the literature. In testicular germ cell cancer, the increased risk to brothers and fathers of patients is estimated to be 8-fold and 4-fold respectively.

Genetic studies of testicular germ cell tumors have revealed recurrent mutations and germline predisposition

Genome wide association studies (GWAS) have been performed in thousands of testicular germ cell tumors and controls, leading to the identification of 44 risk loci [14](Figure 3A), in contrast to ovarian germ cell tumor where only recurrent somatic mutations have been reported (Figure 3B). Genes associated with the 44 risk loci imply three general areas contributing to genetic risk of testicular germ cell tumors: 1) developmental transcriptional regulation as demonstrated by downregulation of the GATA4 transcription factor and specifically variants in PRDM14 and DMR1, both transcriptional regulators of germ cell specification, 2) microtubule or chromosomal assembly including TEX14, a regulator of kinetochore-microtubule assembly in testicular germ cells, and 3) KIT-MAPK signaling, adding to the known importance of the KIT receptor tyrosine kinase that facilitates germ cell migration in normal embryos. In contrast, genetic association

studies in ovarian germ cell tumors are limited to anecdotal studies of small numbers of families that appear to have increased risk.

A study of 137 primary testicular germ cell tumors found the highly recurrent aberration isochrome 12p, or i(12p), in 87% of tumors including both seminoma and non-seminoma. Additionally, authors found a very low mutation frequency of 0.5 mutations/Mb of targeted DNA, placing the mutation rate above pediatric tumors but below most adult tumors[4]. The most common point mutation in testicular germ cell tumors is in the KIT gene, found in ~18% of samples, followed by KRAS and NRAS, at 14% and 4% respectively. Additionally, KRAS and MDM2 gene level amplifications were significantly recurrent. In samples where KIT and KRAS or NRAS mutations co-occurred, KIT mutations were inferred to occur earlier in tumor development.

Ovarian teratomas are defined by presence of tissue from multiple germ cell layers

Mature teratomas, also known as dermoid cysts, are defined by their composition of mature tissues from at least two of the germ layers (ectoderm, endoderm, mesoderm). Mature teratomas are estimated to account for 20% of all ovarian neoplasms and typically present with abdominal pain or a palpable abdominal mass measuring 5-10cm in diameter. In this report, immature teratoma and mixed germ cell tumors containing majority immature components will be the focus. Immature teratoma is loosely defined by the WHO as containing variable amounts of immature embryonic tissue and can exist as part of a pure teratoma or a mixed germ cell tumor. Due to immature teratomas rarity and presence in a mixed germ cell tumor, it is rarely studied in isolation. The argument to study "pure" immature teratoma as an individual condition began in 1976 with a landmark study by Norris and colleagues where age, stage, grade, size, histopathology, and survival metrics were reported for the first time as a distinct clinical entity from mature teratoma and dysgerminoma, where a 81%, 60%, and 30% survival rates were reported for grade 1,2, and 3

respectively [15]. A subsequent retroactive study of 1332 patients with ovarian malignancy found that 10% of cases were germ cell tumors, and 20% of those malignant germ cell tumors were diagnosed immature teratoma, comprising 27/1332 or 2% of all ovarian malignancy [16]. The mean age at diagnosis was 27 years and the main symptom was rapid increased abdominal girth. The abundance of bilateral disease in teratoma patients is ~10% which implies that genetic predisposition may be present in at least some cases.

Gliomatosis peritoneii is a rare condition usually associated with immature ovarian teratoma where mature glial tissue is found in the peritoneum [17]. Gliomatosis peritoneii is not associated with worse outcomes, but some cases developing into high grade glioma have been reported. Prior to the work presented here, the origin of glial cell neoplasms in the peritoneum has been debated and three theories have been proposed: 1) dissemination of malignant cells from immature teratoma in the ovary to the peritoneum followed by epigenetic reprogramming into glia, 2) stem cells from the peritoneum developing into glial cells, and 3) subperitoneal mesenchymal cells. One genetic study found that chromosomal markers heterozygous in normal tissue but homozygous in ovarian teratoma samples were heterozygous in the associated gliomatosis peritoneii tissue, thus theory one from above was ruled out, however this study relied on small numbers of chromosomal markers and has not been reproduced [18].

Mutations in germ cells are referred to as germline mutations and will be present in all cells of the offspring, however we will now refer to mutations in germ cell tumors that contribute to tumor development as somatic mutations because once the tumor is formed the germ cell is restricted to the soma and loses potential to form viable offspring. Unlike their male counterpart, ovarian germ cell tumors do not display recurrent 12p gains [10]. The only observed recurrently mutated gene in ovarian germ cell tumors is KIT, which was reported in 11/41 dysgerminoma samples in one

study, however there is no reliable molecular biomarker to identify ovarian germ cell tumor or distinguish between molecular subtypes [19]. Studies of malignant ovarian germ cell tumor subtypes have been extremely limited and to date no deep sequencing study with power to detect recurrent mutations in non-dysgerminoma subtypes has been performed. Similarly, expression profiling of malignant ovarian germ cell tumors have been limited to small cohorts and have not yielded correlations with clinical significance, or reproducible expression subtypes. One study of male and female germ cell tumors found overexpression of microRNA 371~373 and 302 in malignant germ cell tumors compared to normal controls, however the molecular implication of this finding remains unknown[20].

The suspected origin of ovarian teratoma

In 1969, the cell type and temporal stage originating ovarian teratomas was unknown and under investigation by David Linder. It had been observed that teratomas may be gonadal or extragonadal, have normal karyotypes, and display immature and mature tissue pathology from various tissue layers, thus the hypothesis that teratomas arise from germ cells after or at the time of meiotic division was tested. The authors reasoned that a tumor arising from a germ cell after the first meiotic division would result in a tumor genome that will be homozygous at many sites that were heterozygous in the host cell. Because some but not all markers observed had undergone LOH, the authors concluded that teratomas resulted from cells that had undergone meiosis I or failed meiosis I and then performed the second meiotic division[21]. Due to the small number of loci sampled per genome, and low number of tumor samples, the origin of ovarian teratoma remained inconclusive. An analysis of 102 benign mature teratomas and 2 immature malignant teratomas in 1990 revealed 95% of teratoma genomes showed normal 46,XX karyotypes[22], the remainder showing various aneuploidy with no discernible recurrent pattern. Chromosome 13

centromeric heteromorphisms were also analyzed in tumor versus normal tissue, and it was found that 65% of tumors exhibited loss of heterozygosity (LOH). Parrington et al. [23] concluded that three mechanisms leading to the observed teratoma genomes were possible, suppression of meiosis I, suppression of meiosis II, or whole genome duplication of a haploid gamete. However, due to a lack of reference genome sequence and the absence of polymerase chain reaction based genetic tools, molecular methods at the time relied on chromosome heteromorphisms and isozyme based analyses which typically allowed for an average of ~2 informative markers per host/tumor genome comparison.

Revisiting meiosis is the first step to understanding the link between LOH patterns in germ cell tumors and their cell of origin. Germ cells are the only cells that undergo meiosis and the result is reduced ploidy and recombination of genetic material [24]. In meiosis I, duplicated homologous chromosomes pair to form a bivalent structure containing 4 chromatids, and chromatids undergo genetic recombination where homologous fragments of maternally and paternally derived chromosomes are exchanged. At metaphase of meiosis I, the bivalents line up on the mitotic spindle and cell division results in separation of the duplicated chromosomes into daughter cells. Meiosis I division results in diploid cells containing duplicated chromosomes that are nearly identical and thus contain an entirely homozygous genotype, except for regions that had previously undergone recombination where heterozygosity may be present[24]. Meiotic division II results in separation of the sister chromatids to produce haploid gamete cells. Thus, if meiosis II does not occur in the cell originating a germ cell tumor, the resulting genome will be diploid, with polymorphic loci that were heterozygous in the host homozygous near the centromere, and heterozygous at distal regions that had undergone recombination. If meiosis I division does not occur, but then meiosis II division does occur where sister chromatids are separated, this will result

in a cell that is diploid and remains heterozygous near the centromere and could be heterozygous or homozygous at distal regions depending on the number and location of chromosomal crossing over events.

Chromosomal crossing over rates vary across the genome, species, and even sexes. At least one crossover event per bivalent is required for efficient chromosome alignment and disjunction. Crossover has significant implications for evolution by creating novel combinations of alleles [25]. One logical evolutionary benefit would that favorable alleles can evolve on either homolog, and by crossing over could eventually end up on the same chromosome, maximizing fitness. Crossing over can also lead to more variance among individuals of a population, at the cost of lower mean fitness as favorable interactions between alleles are broken apart by recombination. Chromosome crossover "hotspots" represent sharp local increases in the recombination frequency, however smaller scale variation in crossover rates may also have significant biological meaning. To date a large-scale analysis of crossover rates and their consequence has not been conducted.

Tumor evolution describes the spatial and temporal order of tumor genesis and development

One of the major challenges in cancer medicine is the phenomenon of intratumoral heterogeneity which describes the significant intercellular variation in genetic and epigenetic qualities that fosters the resistant growth and "evolution" of tumors. Inspired by Darwin's principles, cancer biologists have posited that tumors contain groups of cells that act analogously to individuals in a natural population that undergoes the process of natural selection [26]. Several modes of evolution are plausible within tumors: under neutral evolution, genetic changes are randomly generated and inherited across lineages to create a diverse population. When a selective pressure is applied, lineages that contain advantageous phenotypes will continue. Parallel evolution describes the independent evolution of similar traits by different individuals sharing a common ancestor.

Gradual evolution describes the process of many small increases in fitness over time, in contrast to the punctuated model where few but large leaps in fitness occur. Cancer cells may gain diversity by the modification of chromosome number, chromosome structure, somatic mutagenesis, and epigenetic changes[26]. Gaining an accurate picture of tumor evolution is a challenge: traditional methods involve genetic sequencing of a single small tumor region that may not capture the full diversity of the tumor. In tumors with sufficient point mutation frequency and incidence of copy number gain events, the relative temporal order of tumor evolution can be mapped[27, 28]. A study of the evolutionary history of 2,658 cancers found that early events were limited to mutations in a constrained set of driver genes and specific copy number gains such as trisomy 7 in glioblastoma, whereas late events were classified by increased genomic instability and more diverse aberrations. Other methods rely on multi-regional sequencing where spatially distinct tumor pieces are analyzed separately, and then compared to build a qualitative or quantitative model of tumor evolution[29]. A large multi-region mutational analysis study in 2012 reported that 63-69% of all somatic mutations in primary renal carcinomas were not detected in every tumor region. The caveat is that this method is more time consuming, costly, and may not be feasible on small tumors or supported by current hospital diagnostic and research protocols. One hypothesis to explain the unique sensitivity of germ cell tumors to platinum-based chemotherapy is that despite extreme morphological heterogeneity, they are genetically homogeneous and do not contain subpopulations of cells that are able to resist therapy. Another plausible mode of germ cell tumor evolution is that subclonal mutations allow development of diverse subpopulations that arise in late stage disease and lead to worse survival outcomes.

Multi-regional exome sequencing to assess somatic mutation, genetic heterogeneity, clonal relationship of disseminated components and bilateral disease, and genetic mechanism of origin

Here, we utilize modern genomics methods and pathology techniques to study the immature teratoma genome(Figure 4). By deep sequencing of the coding regions of ten pure or mixed majority immature teratomas, we will test the hypothesis that similar to testicular non-seminoma, KIT and KRAS mutations are recurrent. To date, the work presented here is the largest analysis of ovarian immature teratoma genomes and the only that interrogates multiple tumor regions. By annotation of multiple regions per tumor sample by a group of expert gynecologic pathologists followed by exome sequencing of purely isolated tumor region samples, we will test the hypothesis that unique subclonal mutations during tumor development are present in tumor regions with distinct morphology. This could shed light on how teratomas develop tissue from 2-3 germ cell layers, while other germ cell tumors do not. By observing which genes are mutated in tumor regions with differing morphology, we will be able to generate hypotheses for future research into the mechanism of specific genes driving the development of germ cell tumors. By studying bilateral and disseminated disease, we will be able to answer the longstanding question of the clonal relationship between bilateral tumors as well as gliomatosis peritoneii and other related growths in peripheral tissues. Lastly, by observing the zygosity of thousands of SNPs in tumor tissue compared to the normal host throughout the exome, we will be able to accurately test the theories of origination of teratoma by various plausible meiotic failure mechanisms.

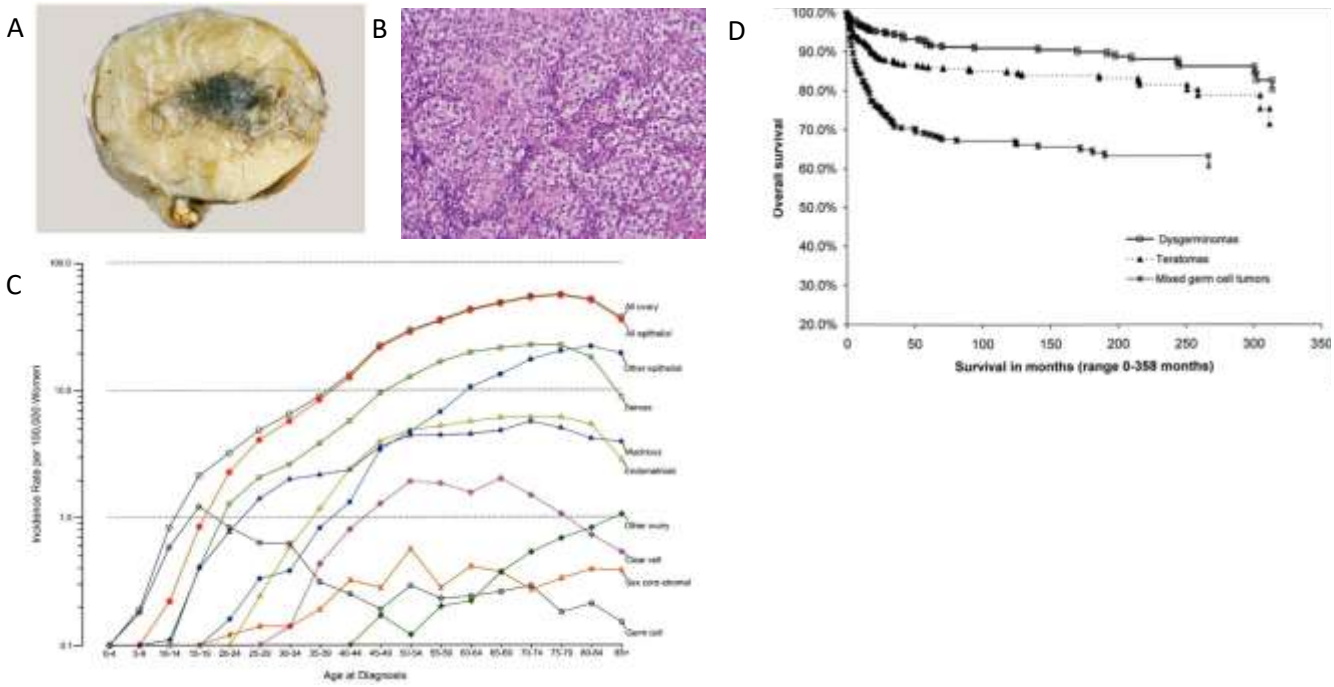


Figure 1.1 Histopathology and epidemiology of ovarian germ cell tumors. A) Example of a cystic mature teratoma with hair. B) Dysgerminoma with cells resembling primordial germ cells. C) Incidence of cancers of the female reproductive tract as a function of age. Incidence of germ cell tumors peaks at the 15-19 age group. D) Survival in months for dysgerminoma, teratoma, and mixed germ cell tumors. Figures adapted from [3, 8, 30, 31].

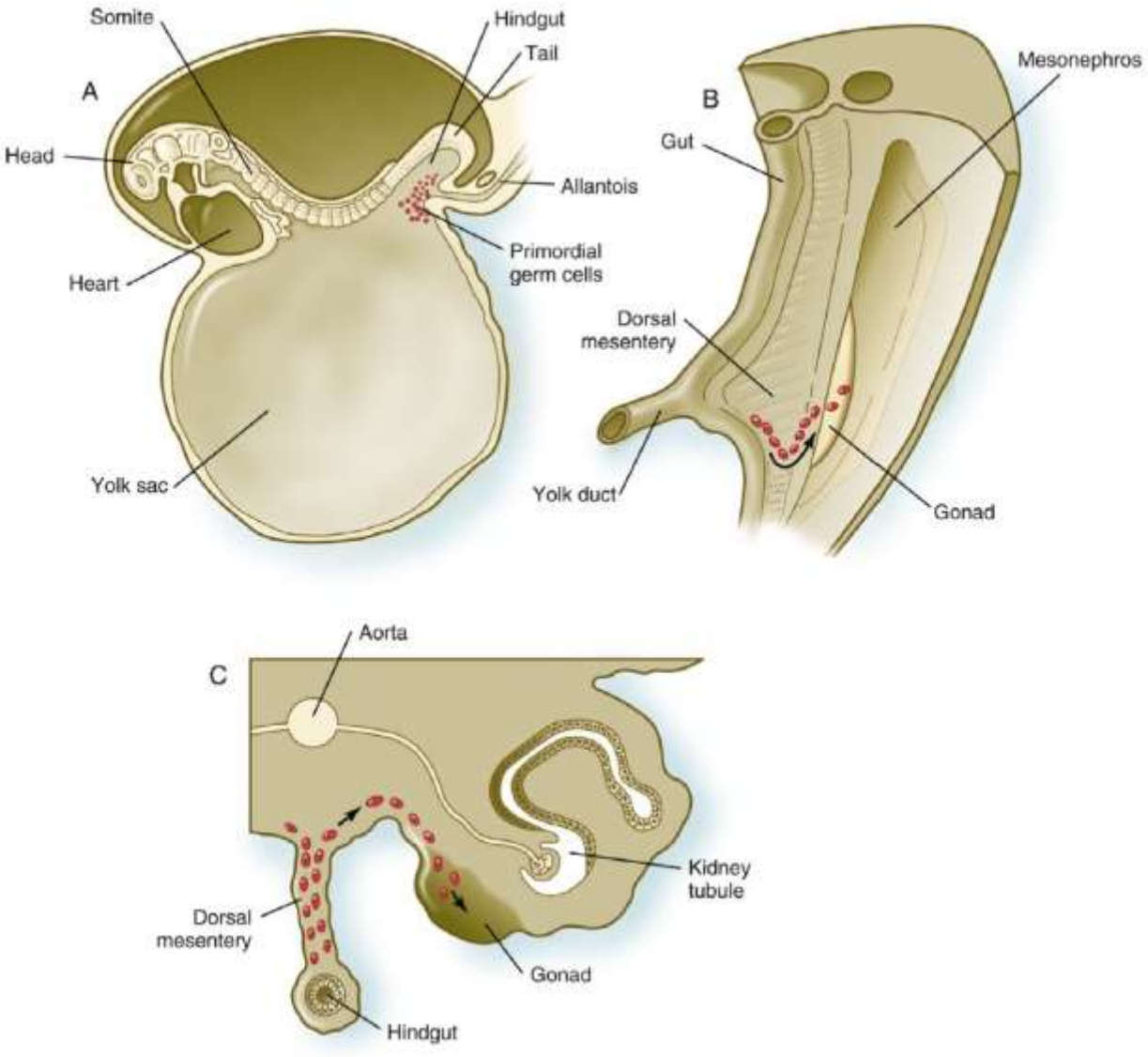


Figure 1.2. Early germ cell development. A) Primordial germ cells (red), the earliest precursors of germ cells, are recognizable at only 24 days after fertilization and are located yolk sac. B-C) Germ cells exit the yolk sac and enter the hindgut epithelium, followed by mitotic reproduction and migration to the gonad. Figure adapted from [32].

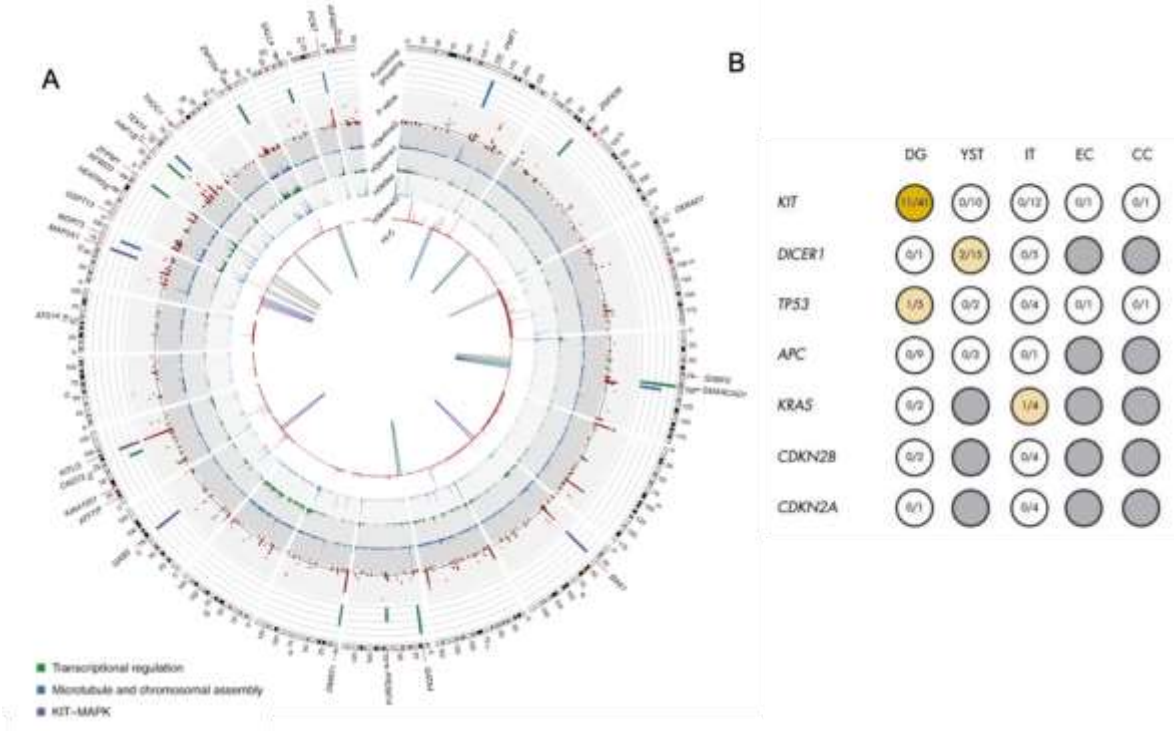


Figure 1.3. Genetics of testicular and ovarian germ cell tumors. A) Circos plot showing all 44 risk loci for testicular germ cell tumors, and their inferred functional associations. B) Summary of somatic mutations detected in ovarian germ cell tumors. Figures adapted from [10, 14]

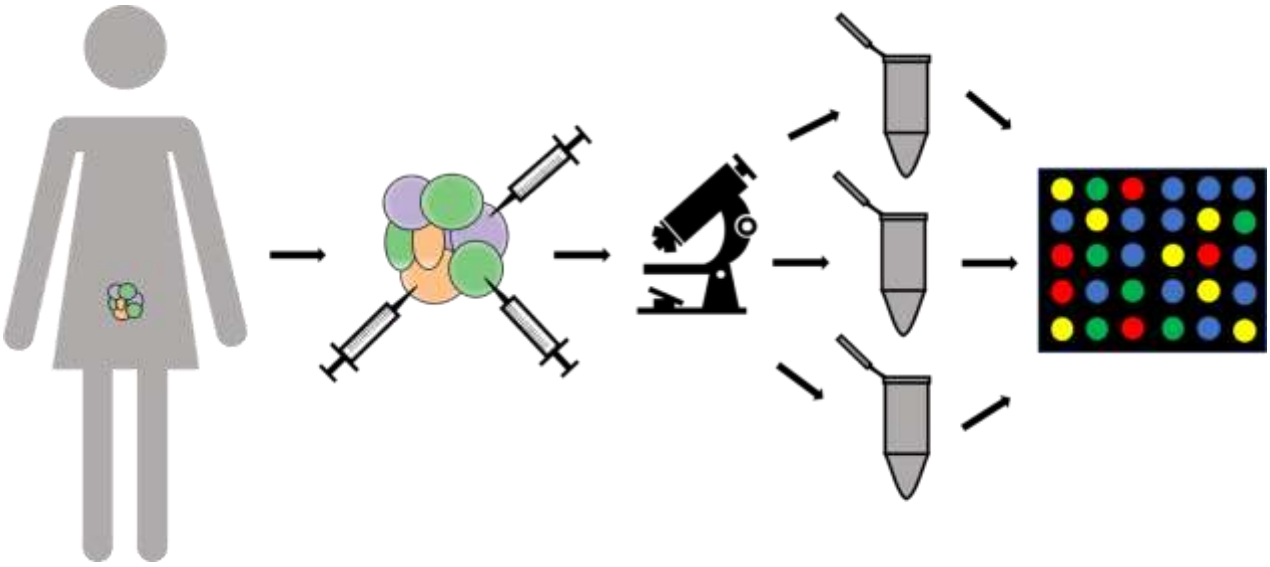


Figure 1.4. Graphical summary of experimental setup. Left to right: Tumors were resected from 10 patients at primary and disseminated regions. Several tumor regions per tumor were resected and annotated by a pathologist. After annotation, exome sequencing was performed on each of 52 tumor regions, including matched normals.

Chapter 1 Work Cited

1. Scultetus, J., *Trichiasis admiranda sive Morbus pilaris mirabilis observatus.* 1658.
2. Damjanov, I., B.B. Knowles, and D. Solter, *The Human Teratomas: Experimental and Clinical Biology*. 2012: Humana Press.
3. Kurman, R.J., *WHO classification of tumours of female reproductive organs.* 2014.
4. Shen, H., et al., *Integrated Molecular Characterization of Testicular Germ Cell Tumors.* Cell Rep, 2018. **23**(11): p. 3392-3406.
5. Williamson, S.R., et al., *The World Health Organization 2016 classification of testicular germ cell tumours: a review and update from the International Society of Urological Pathology Testis Consultation Panel.* Histopathology, 2017. **70**(3): p. 335-346.
6. Richardson, B.E. and R. Lehmann, *Mechanisms guiding primordial germ cell migration: strategies from different organisms.* Nat Rev Mol Cell Biol, 2010. **11**(1): p. 37-49.
7. Gilbert, S.F., *Developmental Biology, 6th Edition*. 2000.
8. Smith, H.O., et al., *Incidence and survival rates for female malignant germ cell tumors.* Obstet Gynecol, 2006. **107**(5): p. 1075-85.
9. Møller, H. and H. Evans, *Epidemiology of gonadal germ cell cancer in males and females*. Apmis, 2003. **111**(1): p. 43-6; discussion 46-8.
10. Kraggerud, S.M., et al., *Molecular characteristics of malignant ovarian germ cell tumors and comparison with testicular counterparts: implications for pathogenesis.* Endocr Rev, 2013. **34**(3): p. 339-76.
11. Dasari, S. and P.B. Tchounwou, *Cisplatin in cancer therapy: molecular mechanisms of action*. Eur J Pharmacol, 2014. **740**: p. 364-78.

12. Solheim, O., et al., *Malignant ovarian germ cell tumors: presentation, survival and second cancer in a population based Norwegian cohort (1953-2009).* Gynecol Oncol, 2013. **131**(2): p. 330-5.

13. Hersmus, R., et al., *The biology of germ cell tumors in disorders of sex development.* Clin Genet, 2017. **91**(2): p. 292-301.

14. Litchfield, K., et al., *Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor.* Nat Genet, 2017. **49**(7): p. 1133-1140.

15. Norris, H.J., H.J. Zirkin, and W.L. Benson, *Immature (malignant) teratoma of the ovary: a clinical and pathologic study of 58 cases.* Cancer, 1976. **37**(5): p. 2359-72.

16. Alwazzan, A.B., et al., *Pure Immature Teratoma of the Ovary in Adults: Thirty-Year Experience of a Single Tertiary Care Center.* Int J Gynecol Cancer, 2015. **25**(9): p. 1616-22.

17. Liang, L., et al., *Gliomatosis peritonei: a clinicopathologic and immunohistochemical study of 21 cases.* Mod Pathol, 2015. **28**(12): p. 1613-20.

18. Snir, O.L., et al., *Frequent homozygosity in both mature and immature ovarian teratomas: a shared genetic basis of tumorigenesis.* Mod Pathol, 2017. **30**(10): p. 1467-1475.

19. Hoei-Hansen, C.E., et al., *Ovarian dysgerminomas are characterised by frequent KIT mutations and abundant expression of pluripotency markers.* Mol Cancer, 2007. **6**: p. 12.

20. Palmer, R.D., et al., *Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets.* Cancer Res, 2010. **70**(7): p. 2911-23.

21. Linder, D., *Gene loss in human teratomas*. Proceedings of the National Academy of Sciences of the United States of America, 1969. **63**(3): p. 699-704.

22. Surti, U., et al., *Genetics and biology of human ovarian teratomas. I. Cytogenetic analysis and mechanism of origin*. American journal of human genetics, 1990. **47**(4): p. 635-643.

23. Parrington, J.M., L.F. West, and S. Povey, *The origin of ovarian teratomas*. Journal of medical genetics, 1984. **21**(1): p. 4-12.

24. Alberts, B.J.A.L., J., *Molecular Biology of the Cell. 4th edition*. 2002.

25. Haenel, Q., et al., *Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics*. Mol Ecol, 2018. **27**(11): p. 2477-2497.

26. McGranahan, N. and C. Swanton, *Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future*. Cell, 2017. **168**(4): p. 613-628.

27. Durinck, S., et al., *Temporal dissection of tumorigenesis in primary cancers*. Cancer Discov, 2011. **1**(2): p. 137-43.

28. Gerstung, M., et al., *The evolutionary history of 2,658 cancers*. Nature, 2020. **578**(7793): p. 122-128.

29. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing*. N Engl J Med, 2012. **366**(10): p. 883-892.

30. Quirk, J.T., N. Natarajan, and C.J. Mettlin, *Age-specific ovarian cancer incidence rate patterns in the United States*. Gynecol Oncol, 2005. **99**(1): p. 248-50.

31. Shaaban, A.M., et al., *Ovarian malignant germ cell tumors: cellular classification and clinical and imaging features*. Radiographics, 2014. **34**(3): p. 777-801.

32. Carlson, B.M., *Gametogenesis*. Reference Module in Biomedical Sciences, 2014.

Chapter 1.4: Multi-region exome sequencing of ovarian teratomas reveals 2N near-diploid genomes, paucity of somatic mutations, and extensive allelic imbalances shared across mature, immature, and disseminated components. Heskett et al 2020.

Multi-region exome sequencing of ovarian teratomas reveals 2N near-diploid genomes, paucity of somatic mutations, and extensive allelic imbalances shared across mature, immature, and disseminated components

Michael B. Heskett[1], John Z. Sanborn[2], Christopher Boniface[1], Benjamin Goode[3], Jocelyn Chapman[4], Karuna Garg[3], Joseph T. Rabban[3], Charles Zaloudek[3], Stephen C. Benz[2], Paul T. Spellman[1], David A. Solomon[3]*, and Raymond J. Cho[5]*

1. Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA

2. NantOmics, LLC, Culver City, CA, USA

3. Department of Pathology, University of California, San Francisco, CA, USA

4. Division of Gynecologic Oncology, Department of Obstetrics, Gynecology & Reproductive Sciences, University of California, San Francisco, CA, USA

5. Department of Dermatology, University of California, San Francisco, CA, USA

*To whom correspondence should be addressed:

David A. Solomon, MD, PhD, Department of Pathology, University of California, San Francisco, 513 Parnassus Ave, Health Sciences West 451, San Francisco, CA 94143, United States, Ph: (415) 514-9761, Email: david.solomon@ucsf.edu

Raymond J. Cho, MD, PhD, Department of Dermatology, University of California, San Francisco, 1701 Divisadero Street, 3rd floor, San Francisco, CA 94115, United States, Ph: (415) 650-5208, Email: raymond.cho@ucsf.edu

Keywords: immature teratoma, mature teratoma, dermoid cyst, ovarian cancer, germ cell tumor, meiosis, molecular pathology

Abstract

Immature teratoma is a subtype of malignant germ cell tumor of the ovary that occurs most commonly in the first three decades of life, frequently with bilateral ovarian disease. Despite being the second most common malignant germ cell tumor of the ovary, little is known about its genetic underpinnings. Here we performed multi-region whole exome sequencing to interrogate the genetic zygosity, clonal relationship, DNA copy number, and mutational status of 52 pathologically distinct tumor components from 10 females with ovarian immature teratomas, with bilateral tumors present in 5 cases and peritoneal dissemination in 7 cases. We found that ovarian immature teratomas are genetically characterized by 2N near-diploid genomes with extensive loss of heterozygosity and an absence of genes harboring recurrent somatic mutations or known oncogenic variants. All components within a single ovarian tumor (immature teratoma, mature teratoma with different histologic patterns of differentiation, and yolk sac tumor) were found to harbor an identical pattern of loss of heterozygosity across the genome, indicating a shared clonal origin. In contrast, the 4 analyzed bilateral teratomas showed distinct patterns of zygosity changes in the right versus left sided tumors, indicating independent clonal origins. All disseminated teratoma components within the peritoneum (including gliomatosis peritonei) shared a clonal pattern of loss of heterozygosity with either the right or left primary ovarian tumor. The observed genomic loss of heterozygosity patterns indicate that diverse meiotic errors contribute to the formation of ovarian immature teratomas, with 11 out of the 15 genetically distinct clones determined to result from the failure of meiosis I or II. Overall, these findings suggest that copy-neutral loss of heterozygosity resulting from meiotic abnormalities may be sufficient to generate ovarian immature teratomas from germ cells.

Introduction

Germ cell tumors (GCTs) are a diverse group of neoplasms that display remarkable heterogeneity in their anatomical site, histopathology, prognosis, and molecular characteristics [1]. GCTs can occur in the ovaries, testes, and extragonadal sites, with the most common extragonadal locations being the anterior mediastinum, retroperitoneum, and intracranially in the pineal region [2]. GCTs are classified by the World Health Organization into seven histological subtypes: mature teratoma, immature teratoma, seminoma/dysgerminoma/germinoma (depending on site of origin in the testis, ovary, or extragonadal), yolk sac tumor, embryonal carcinoma, choriocarcinoma, and mixed germ cell tumor [3].

GCTs are the most common non-epithelial tumors of the ovary, but only account for approximately 3% of all ovarian cancers [4]. Approximately 90% of ovarian GCTs are composed entirely of mature teratoma (commonly termed “dermoid cyst”), which is the only benign subtype of ovarian GCT. Among the malignant subtypes, dysgerminoma is the most common and immature teratoma is the second most common. Ovarian teratomas contain tissue elements from at least 2 of the 3 germ cell layers and frequently display a disorganized mixture of mature tissues including skin and hair (ectoderm), neural tissue (ectoderm), fat (mesoderm), muscle (mesoderm), cartilage (mesoderm), bone (mesoderm), respiratory epithelium (endoderm), and gastrointestinal epithelium (endoderm). Teratomas can occur in the mature form, composed exclusively of mature tissues, or the immature form, which contains variable amounts of immature elements (usually primitive neuroectodermal tissue consisting of primitive neural tubules) in a background of mature teratoma [5]. Not infrequently, malignant GCTs of the ovary contain a mixture of different histologic subtypes (*e.g.* both dysgerminoma and yolk sac tumor), for which the designation mixed germ cell tumor is used, often with the approximate fraction of each histologic subtype specified by the diagnostic pathologist. Extensive tissue sampling and microscopic review of ovarian GCTs are required to appropriately evaluate for the presence of admixed malignant subtypes, which is critical for appropriately guiding prognosis and patient management.

The majority of ovarian GCTs (57%) are confined to the ovary at time of diagnosis (stage I) which confers a 99% 5-year survival [4]. Even when distant metastases are present at time of diagnosis (stage IV), 5-year survival of ovarian GCT is relatively high at 69% [4]. This long-term survival in females even with disseminated or metastatic ovarian GCTs reflects the sensitivity of these tumors to the standard cytotoxic chemotherapy regimen of bleomycin, etoposide, and cisplatin [1].

Somatic mutation and DNA copy number analysis of testicular GCTs has now been performed by The Cancer Genome Atlas Research Network and several other groups [6-14]. These analyses have revealed a very low mutation rate (approximately 0.3 somatic mutations per Mb) and only three genes harboring recurrent somatic mutations at significant frequency (*KIT*, *KRAS*, and *NRAS*), in which mutations are exclusively present in seminomas but not non-seminomatous GCTs [7-14]. Copy number analysis has revealed that testicular GCTs are often hyperdiploid, with the majority (>80%) harboring isochromosome 12p or polysomy 12p that is present in both seminomas and non-seminomatous GCTs [6, 9, 12, 13, 14]. Similar oncogenic *KIT* and *KRAS* mutations as well as polysomy 12p have also been frequently found in ovarian dysgerminomas and intracranial germinomas, indicating a shared molecular pathogenesis with testicular seminomas [15-20].

Beyond dysgerminomas, few studies have performed genome-level analysis of ovarian GCTs, and the genetic basis of ovarian teratomas (both mature and immature forms) remains unknown. Polysomy 12 and *KIT* mutations have been found in ovarian mixed germ cell tumors containing a dysgerminoma component, but have not been identified in pure teratomas [20]. Early studies of ovarian mature teratomas reported that tumor karyotypes were nearly always normal (*i.e.* 46,XX), but chromosomal zygosity markers were often homozygous in the tumor [21-24]. This loss of heterozygosity may be explained by the hypothesis that teratomas and other germ cell tumors arise from primordial germ cells due to one of five different plausible meiotic abnormalities, each producing distinct chromosomal patterns of homozygosity [23-26]. Parthenogenesis (from the Greek *parthenos*: 'virgin', and *genesis*: 'creation') is used to describe the development of germ cell tumors from unfertilized germ cells via these different mechanisms of origin, which potentially include failure of meiosis I, failure of meiosis II,

whole genome duplication of a mature ovum, and fusion of two ova. However, no studies to date have used genome-level sequencing analysis to identify the specific parthenogenetic mechanism giving rise to individual ovarian GCTs.

Here we present the results of multi-region whole exome sequencing of 52 pathologically distinct tumor components from 10 females with ovarian immature teratomas, with bilateral tumors present in 5 cases and peritoneal dissemination in 7 cases. Our analyses define ovarian immature teratoma as a genetically distinct entity amongst the broad spectrum of human cancer types studied to date, which is characterized by a 2N near-diploid genome, paucity of somatic mutations, and extensive allelic imbalances. Our results further shed light on the parthenogenetic origin of ovarian teratomas and reveal that diverse meiotic errors are likely to drive development of this germ cell tumor.

Materials and Methods

Study population and tumor specimens

This study was approved by the Institutional Review Board of the University of California, San Francisco. Ten patients who underwent resection of ovarian immature teratomas at the University of California, San Francisco Medical Center between the years 2002-2015 were included in this study. All tumor specimens were fixed in 10% neutral-buffered formalin and embedded in paraffin. Pathologic review of all tumor specimens was performed to confirm the diagnosis by a group of expert gynecologic pathologists (K.G., J.T.R., C.Z., and D.A.S.).

Whole exome sequencing

Tumor tissue from each of the indicated ovarian and disseminated germ cell tumor components was selectively punched from formalin-fixed, paraffin-embedded blocks using 2.0 mm disposable biopsy punches (Integra Miltex Instruments, cat# 33-31-P/25). These punches were made into areas histologically visualized to be composed entirely of the indicated germ cell component (*e.g.* immature teratoma, mature teratoma, yolk sac tumor, gliomatosis peritonei). Uninvolved normal fallopian tube was also selectively punched from formalin-fixed, paraffin-embedded blocks as a source of constitutional DNA for each of the ten patients. Genomic DNA was extracted from these tumor and matched normal tissue samples using the QIAamp DNA FFPE Tissue Kit (Qiagen) according to the manufacturer's protocol. 500 ng of genomic DNA was used as input for capture employing the xGen Exome Research Panel v1.0 (Integrated DNA Technologies). Hybrid-capture libraries were sequenced on an Illumina HiSeq 4000 instrument.

Mutation calling and loss of heterozygosity analysis

Sequence reads were aligned to the hg19 reference genome using Burrows-Wheeler Alignment tool [27]. Duplicate reads were removed and base quality scores recalibrated with GATK prior to downstream analysis [28]. Candidate somatic mutations were identified with MuTect v1.1.5 with the minimum mapping quality parameter set to 20.

dbSNP build 150 was used to identify and remove SNPs. The following additional filters were applied to candidate mutations from MuTect output: minimum tumor depth 30, minimum normal depth 15, minimum variant allele frequency 15%, maximum variant allele presence in normal 2%. Finally, all candidate mutations were manually reviewed in the Integrative Genome Viewer to remove spurious variant calls likely arising from sequencing artifact [29, 30]. FACETS was used to determine allele-specific copy number and loss of heterozygosity regions across the genome [31]. To determine genetic mechanism of origin, tumors were classified into one of five plausible categories based on the zygosity status at centromeric and distal regions, as described by Surti et al [25]. For visualization of zygosity changes across the genome in the tumor specimens, the absolute difference between theoretical heterozygosity (allele frequency = 0.5) of tumor versus normal was plotted.

Results

Patient cohort

Clinical data from the patient cohort is summarized in Table 1. The 10 females ranged in age at time of initial surgery from 8-29 years (median 17 years). None were known to have Turner syndrome or other gonadal dysgenesis disorder, nor any known familial tumor predisposition syndrome. All patients underwent resection of a primary ovarian mass, along with debulking of disseminated disease observed in the peritoneum at time of initial oophorectomy for 5 patients. Bilateral ovarian tumors were present in 5 of the 10 patients, 2 with synchronous disease at time of initial diagnosis (a and b) and 3 with metachronous disease that was identified and resected during the period of clinical follow-up (g, h, and j). Primary ovarian tumor size ranged from 4-30 cm (median 15 cm). Four of the patients were treated with adjuvant chemotherapy using bleomycin, etoposide, and cisplatin after initial surgery based on the presence of disseminated immature teratoma in the peritoneum (a, b, e, and k). A fifth patient was treated with adjuvant chemotherapy using bleomycin, etoposide, and cisplatin following resection of a synchronous ovarian immature teratoma at 4.8 years after resection of a contralateral ovarian mature teratoma (g). One exceptional 14-year-old patient (d) initially underwent resection of a unilateral 18 cm ovarian immature teratoma and debulking of disseminated peritoneal disease. Subsequent PET/CT showed widespread bulky lymphadenopathy. She underwent resection of a supraclavicular lymph node at 0.6 years after initial oophorectomy which contained metastatic primitive neuroectodermal tumor (PNET) and atypical gliomatosis histologically resembling an anaplastic astrocytoma of the central nervous system. She was treated with intensive multiagent chemotherapy including vincristine, doxorubicin, cyclophosphamide, ifosfamide, and etoposide. Over the next three years, she underwent additional resections of recurrent/progressive disease in the peritoneum, cyberknife radiotherapy to left axilla, and multiple courses of chemotherapy, first with temozolomide and then with cyclophosphamide and topotecan. She remains alive with stable disease at last clinical follow-up (6.6 years after initial surgery). All other patients in this cohort also remain alive with stable disease or without evidence of disease recurrence at last clinical follow-up (range 2.4-15.3 years, median 6.6 years, excluding patient i with no clinical follow-up data after initial resection).

Histologic features of the ovarian immature teratomas

Pathologic diagnosis for the ovarian germ cell tumors is summarized in Table 1, and representative photomicrographs are shown in Figure 1. All 10 patients had primary ovarian immature teratomas composed of primitive neural tubules in a background of mature teratoma. In 2 patients, there were additionally admixed small foci of yolk sac tumor and embryonal carcinoma, thereby warranting designation as mixed germ cell tumor, although mature and immature teratoma were the predominant elements in both cases. Five patients also had teratomas involving the contralateral ovary, 2 of which were synchronous and 3 of which were metachronous. The contralateral ovarian tumors were also immature teratomas in 2 patients (a and h), whereas the contralateral ovarian tumors were composed entirely of mature teratoma in 3 patients (b, g, and j). Disseminated disease was found in the peritoneum of 7 patients, which consisted of a combination of immature and mature elements in 5 patients and mature elements only in 2 patients. The disseminated immature elements in one of these patients (d) was histologically diagnosed as primitive neuroectodermal tumor (PNET), as it was composed of sheets of primitive small round blue cells with diffuse immunoreactivity for synaptophysin and without organization into neural tubules or evidence of neuroglial differentiation. Six patients had peritoneal implants composed of mature glial tissue that has been termed gliomatosis peritonei. This gliomatosis peritonei was of low cellularity and composed of cytologically bland glial cells in 5 patients, whereas the gliomatosis peritonei was hypercellular and composed of cytologically atypical glial cells resembling anaplastic astrocytoma of the central nervous system in 1 patient (d).

Multi-region whole exome sequencing of ovarian immature teratomas

Genomic DNA was extracted from 52 tumor regions consisting of ovarian immature teratoma, mature teratoma, yolk sac tumor, and disseminated teratomatous elements, along with uninvolved normal fallopian tube tissue from the 10 female patients (Table 2). Hybrid exome capture and massively parallel sequencing by synthesis on an Illumina platform was performed to an average depth of 203x per sample, as described in the

Methods. Sequencing metrics are displayed in Supplementary Table 1. The number of tumor regions sequenced per patient ranged from 2 to 9, with a median of 4.

Paucity of somatic single nucleotide variants in ovarian immature teratomas

Based on this whole exome sequencing of 52 tumor samples, we identified a total of only 31 unique high-confidence somatic nonsynonymous mutations (Supplementary Table 2). Despite high sequencing depth, we detected somatic nonsynonymous mutations in only 21 of the 52 samples, and the average number of somatic nonsynonymous mutations in the mutated samples was 0.8 per exome. The mean somatic mutation burden (commonly also referred to as total mutation burden or TMB) per tumor sample was 0.02 non-synonymous mutations per Mb, which is among the lowest of any human cancer type that has been analyzed to date.

Only 1 of the somatic nonsynonymous mutations (*PKFP* p.A158V in patient a, RefSeq transcript NM_002627) was present in all tumor regions sequenced from a single patient, thereby indicating its clonality and acquisition early during tumorigenesis. However, the other 30 somatic nonsynonymous mutations were present only in a single tumor region or a subset of the tumor regions sequenced, thereby indicating their subclonality and acquisition later during tumorigenesis. For example, the *PKFP* p.A158V mutation was present in all tumor regions sequenced from patient a, including the immature and both mature teratoma components from the left ovary, as well as the disseminated immature teratoma, mature teratoma, and yolk sac tumor components in the peritoneum. In contrast, the *CCS* p.R112C (RefSeq transcript NM_005125) mutation was exclusively present in the ovarian mature teratoma component with neuroglial differentiation, and the *CIITA* p.R2C (RefSeq transcript NM_000246) mutation was only present in the disseminated immature teratoma and yolk sac tumor components. Thus, none of these 30 somatic nonsynonymous mutations could have plausibly been the initiating genetic driver in this cohort of ovarian immature teratomas.

No genes were identified to harbor recurrent somatic nonsynonymous mutations across the 10 patients (*i.e.* no gene was mutated in more than a single patient). Furthermore, no well-described oncogenic variants (e.g. *BRAF* p.V600E) were identified

in any of the 52 tumors samples. Among the 723 genes currently annotated in the Cancer Gene Census of the Catalog of Somatic Mutations in Cancer (COSMIC) database version 90 release, only 4 were identified to harbor somatic nonsynonymous mutations in this ovarian immature teratoma cohort. However, the variants in these 4 genes (*TP53*, *NF1*, *CTNNB1*, and *NOTCH2*) were each found in a single tumor sample in this cohort, were all non-truncating missense variants, and are not known recurrent somatic mutations in the current version of the COSMIC database. Thus, the functional significance of the identified mutations in these 4 genes is uncertain, and they may likely represent bystander alterations rather than driver mutations. Although *KIT*, *KRAS*, *NRAS*, and *RRAS2* are recurrently mutated oncogenes that drive ovarian dysgerminomas and testicular germ cell tumors [14, 19, 20], we found no mutations in these genes in this cohort of ovarian immature teratomas.

Ovarian immature teratomas have 2N diploid or near-diploid genomes with extensive loss of heterozygosity

Using FACETS to infer copy number status and the genotype data of common polymorphisms from the exome sequencing, we next assessed the chromosomal copy number and zygosity status of the 52 tumor samples (Table 2). All of the 52 tumor samples were found to harbor 2N diploid or near-diploid genomes. All tumor samples from 6 of the patients had normal 46,XX diploid genomes. All tumor samples from 3 of patients had near-diploid genomes with clonal gain of a single whole chromosome (+3 in patient d, +14 in patient i, and +10 in patient k). In patient b with bilateral ovarian teratomas, the mature teratoma from the right ovary harbored a normal 46,XX diploid genome, whereas all tumor samples from the left ovary and all disseminated peritoneal tumor samples harbored near-diploid genomes with clonal gains of whole chromosomes 3 and X. No focal amplification or deletion events were identified in any of the 52 tumor samples. None of the tumor samples harbored isochromosome 12p or polysomy 12p.

We next plotted the absolute change in allele frequency (ΔAF) for the 52 tumor samples based on the genotype of common polymorphisms across each of the chromosomes, using an average of approximately 17,000 informative loci per genome.

Whereas an allele frequency of 0.5 equals the normal heterozygous state for a diploid genome, an allele frequency of 0.0 or 1.0 equals a homozygous state, which could be due to either chromosomal copy loss or copy-neutral loss of heterozygosity. We observed extensive copy-neutral loss of heterozygosity across the genomes of each of the 52 tumor samples from all 10 patients (Figure 2).

Identical patterns of genomic loss of heterozygosity among mature, immature, and disseminated components in an ovarian teratoma confirm a single clonal origin

We next compared the regions of the genome affected by copy-neutral loss of heterozygosity among the different tumor regions sequenced for each individual patient. In the 5 females with unilateral ovarian disease (patients c, d, e, i, and k), we observed the identical pattern of allelic imbalance across the genome in each of the different tumor components, including immature teratoma, mature teratoma with different histologic patterns of differentiation, and disseminated teratomatous elements in the peritoneum. These results confirm a single clonal origin for all teratomatous components, both in the primary ovarian tumor and disseminated in the peritoneum, for women with unilateral ovarian immature teratomas.

Bilateral ovarian teratomas originate independently

Four patients in this cohort (b, g, h, and j) had bilateral ovarian teratomas that were both independently sequenced and analyzed for patterns of copy-neutral loss of heterozygosity across the genome. We found that tumors from the left and right ovaries had different patterns of allelic imbalance across the genome in each of the different tumor components studied, providing evidence that bilateral ovarian teratomas originate independently. Furthermore, all of the peritoneal disseminated components harbored a pattern of allelic imbalance that was identical to one of the two ovarian tumors, enabling deduction of the specific ovarian tumor from which the disseminated disease was clonally related. For example, patient h is an 8-year-old girl who initially underwent resection of a 17 cm immature teratoma from the left ovary, and then 9 years later underwent resection

of a 16 cm immature teratoma from the right ovary as well as debulking of disseminated disease in the peritoneum (gliomatosis peritonei). The immature teratoma and two mature teratoma regions studied from the left ovary had an identical pattern of allelic imbalance, whereas the immature teratoma and two mature teratoma regions studied from the right ovary had an identical pattern of allelic imbalance that was distinct from the tumor elements in the contralateral ovary. Additionally, the gliomatosis peritonei had an identical pattern of allelic imbalance as the immature and mature teratoma components from the right ovary (Figure 3).

Patterns of genomic loss of heterozygosity in ovarian immature teratomas can be used to deduce meiotic error mechanism of origin

Five distinct parthenogenetic mechanisms of origin have been proposed to describe the development of germ cell tumors from unfertilized germ cells, which include failure of meiosis I, failure of meiosis II, whole genome duplication of a mature ovum, and fusion of two ova. Distinct chromosomal zygosity patterns are predicted to result from each of these different mechanisms [25], which are illustrated in Figure 4. We used the chromosomal zygosity patterns from the whole exome sequencing data to deduce the meiotic mechanism of origin for the 15 distinct tumor clones identified in the 10 female patients. Five of the tumor clones were deduced to result from failure of meiosis I, 6 from failure of meiosis II, 3 from whole genome duplication of a mature ovum, and 1 from fusion of two ova (Table 2). These findings indicate that failure at multiple stages during germ cell development can contribute to the development of ovarian teratomas.

Discussion

We present the first multi-region exome sequencing analysis of ovarian immature teratomas including mature, immature, and disseminated components. We report a strikingly low abundance of somatic mutations and infrequent copy number aberrations, without pathogenic mutations identified in any well-described oncogenes or tumor suppressor genes, as well as an absence of any novel genes harboring recurrent mutations across the cohort. We generated high-resolution zygosity maps of ovarian teratomas that deepen understanding of the parthenogenetic mechanisms of origin of ovarian teratomas from primordial germ cells originally proposed nearly 50 years ago [21]. Ovarian teratoma is genetically unique among all human tumor types studied to date given its extremely low mutation rate and extensive genomic loss of heterozygosity. Our findings suggest that meiotic non-disjunction events producing a 2N near-diploid genome with extensive allelic imbalances are responsible for the development of ovarian immature teratomas.

Analysis of the multi-region exome sequencing data was used to study the clonal relationship of immature and mature teratoma elements, as well as admixed foci of yolk sac tumor, and also disseminated teratoma in the peritoneum. We find that all these different tumor components are indistinguishable based on chromosomal copy number alterations and loss of heterozygosity patterns, indicating a shared clonal origin. This finding suggests that epigenetic differences are likely responsible for the striking variation in differentiation patterns in teratomas.

Notably, gliomatosis peritonei is a rare phenomenon in which deposits of mature glial tissue are found in the peritoneum, which occurs almost exclusively in association with immature teratoma of the gonads [32, 33]. Two theories currently exist to explain the origin of gliomatosis peritonei: the first being that it is derived from peritoneal dissemination of teratoma with differentiation into mature glial cells, and the other being spontaneous metaplasia of peritoneal stem cells to glial tissue in individuals with gonadal teratoma [34, 35]. A prior study of five samples had concluded that gliomatosis peritonei was genetically unrelated to the primary ovarian teratoma based on zygosity analysis of a small number of microsatellite markers [35]. However, our study based on genotyping

data from thousands of informative polymorphic loci unequivocally demonstrates that gliomatosis peritonei is clonally related to the ovarian primary immature teratoma in all cases, thereby confirming the first theory of origin.

While all ovarian and disseminated tumor components in the 5 patients with unilateral disease in this cohort were found to be clonally related, 4 patients had bilateral ovarian teratomas that were independently analyzed and found to have distinct clonal origins. We found that tumors from the left and right ovaries had different patterns of loss of heterozygosity across the genome in each of the different tumor components that were sequenced, providing evidence that bilateral ovarian teratomas originate independently. Additionally, all of the disseminated components in the peritoneum harbored a pattern of allelic imbalance that was identical to one of the two ovarian tumors, enabling assignment of origin to the specific ovarian primary tumor. Why a significant proportion of women with ovarian teratomas also develop genetically independent teratomas in the contralateral ovary (either synchronously or metachronously) remains undefined. Analysis of the constitutional DNA sequence data from the 10 patients in our cohort, 5 of whom had bilateral ovarian teratomas, did not identify pathogenic variants in the germline known to be associated with increased cancer risk. However, the possibility of an unidentified germline risk allele(s) responsible for teratoma development remains a possibility. Given the extensive loss of heterozygosity across the genomes of ovarian teratomas, pinpointing any single responsible gene amongst the numerous common regions of allelic imbalance is a significant obstacle.

In summary, our multi-region whole exome sequencing analysis of ovarian immature teratomas has revealed that multiple different meiotic errors can give rise to these genetically distinct tumors that are characterized by extensive allelic imbalances and a paucity of somatic mutations and copy number alterations.

Acknowledgements

This study was supported by NIH Director's Early Independence Award (DP5 OD021403) and the UCSF Physician-Scientist Scholar Program to D.A.S.

Conflict of interest

The authors declare that they have no conflicts of interest to disclose.

References

1. Pectasides D, Pectasides E, Kassanos D. Germ cell tumors of the ovary. Cancer Treat Rev. 2008;34:427-41.

2. McKenney JK, Heerema-McKenney A, Rouse RV. Extragonadal germ cell tumors: a review with emphasis on pathologic features, clinical prognostic variables, and differential diagnostic considerations. Adv Anat Pathol. 2007;14:69-92.

3. Kurman RJ, Carcangiu ML, Herrington CS, Young RH. World Health Organization Classification of Tumours of Female Reproductive Organs. 4th ed. Lyon: WHO Press, 2014.

4. Torre LA, Trabert B, DeSantis CE, Miller KD, Samimi G, Runowicz CD, et al. Ovarian cancer statistics, 2018. CA Cancer J Clin. 2018;68:284-96.

5. Jorge S, Jones NL, Chen L, Hou JY, Tergas AI, Burke WM, et al. Characteristics, treatment and outcomes of women with immature ovarian teratoma, 1998-2012. Gynecol Oncol. 2016;142:261-6.

6. Gibas Z, Prout GR, Pontes JE, Sandberg AA. Chromosome changes in germ cell tumors of the testis. Cancer Genet Cytogenet. 1986;19:245-52.

7. Mulder MP, Keijzer W, Verkerk A, Boot AJ, Prins ME, Splinter TA, et al. Activated ras genes in human seminoma: evidence for tumor heterogeneity. Oncogene. 1989;4:1345-51.

8. Tian Q, Frierson HF Jr, Krystal GW, Moskaluk CA. Activating c-kit gene mutations in human germ cell tumors. Am J Pathol. 1999;154:1643-7.

9. Roelofs H, Mostert MC, Pompe K, Zafarana G, van Oorschot M, van Gurp RJ, et al. Restricted 12p amplification and RAS mutation in human germ cell tumors of the adult testis. Am J Pathol. 2000;157:1155-66.

10. Kemmer K, Corless CL, Fletcher JA, McGreevey L, Haley A, Griffith D, et al. KIT mutations are common in testicular seminomas. Am J Pathol. 2004;164:305-13.

11. Coffey J, Linger R, Pugh J, Dudakia D, Sokal M, Easton DF, et al. Somatic KIT mutations occur predominantly in seminoma germ cell tumors and are not predictive of bilateral disease: report of 220 tumors and review of literature. Genes Chromosomes Cancer. 2008;47:34-42.

12. Litchfield K, Summersgill B, Yost S, Sultana R, Labreche K, Dudakia D, et al. Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nat Commun. 2015;6:5973.

13. Cutcutache I, Suzuki Y, Tan IB, Ramgopal S, Zhang S, Ramnarayanan K, et al. Exome-wide sequencing shows low mutation rates and identifies novel mutated genes in seminomas. Eur Urol. 2015;68:77-83.

14. Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, et al. Integrated molecular characterization of testicular germ cell tumors. Cell Rep. 2018;23:3392-406.

15. Riopel MA, Spellerberg A, Griffin CA, Perlman EJ. Genetic analysis of ovarian germ cell tumors by comparative genomic hybridization. Cancer Res. 1998;58:3105-10.

16. Cossu-Rocca P, Zhang S, Roth LM, Eble JN, Zheng W, Karim FW, et al. Chromosome 12p abnormalities in dysgerminoma of the ovary: a FISH analysis. Mod Pathol. 2006;19:611-5.

17. Kraggerud, SM, Hoei-Hansen CE, Alagaratnam S, Skotheim RI, Abeler VM, Rajpert-De Meyts E, et al. Molecular characteristics of malignant ovarian germ cell tumors and comparison with testicular counterparts: implications for pathogenesis. Endocr Rev. 2013;34:339-76.

18. Fukushima S, Otsuka A, Suzuki T, Yanagisawa T, Mishima K, Mukasa A, et al. Mutually exclusive mutations of KIT and RAS are associated with KIT mRNA expression and chromosomal instability in primary intracranial pure germinomas. Acta Neuropathol. 2014;127:911-25.

19. Schulte SL, Waha A, Steiger B, Denkhaus D, Dörner E, Calaminus G, et al. CNS germinomas are characterized by global demethylation, chromosomal instability and mutational activation of the Kit-, Ras/Raf/Erk- and Akt-pathways. Oncotarget. 2016;7:55026-42.

20. Van Nieuwenhuysen E, Busschaert P, Neven P, Han SN, Moerman P, Liontos M, et al. The genetic landscape of 87 ovarian germ cell tumors. Gynecol Oncol. 2018;151:61-8.

21. Linder D. Gene loss in human teratomas. Proc Natl Acad Sci USA. 1969;63:699-704.

22. Carritt B, Parrington JM, Welch HM, Povey S. Diverse origins of multiple ovarian teratomas in a single individual. Proc Natl Acad Sci USA. 1982;79:7400-4.

23. Parrington JM, West LF, Povey S. The origin of ovarian teratomas. J Med Genet. 1984;21:4-12.

24. Ohama K, Nomura K, Okamoto E, Fukuda Y, Ihara T, Fujiwara A. Origin of immature teratoma of the ovary. Am J Obstet Gynecol. 1985;152:896-900.

25. Surti U, Hoffner L, Chakravarti A, Ferrell RE. Genetics and biology of human ovarian teratomas. I. Cytogenetic analysis and mechanism of origin. Am J Hum Genet. 1990;47:635-43.

Figure Legends

Fig. 1 Histology images of the ovarian immature teratomas from three representative patients that were studied by whole exome sequencing. Shown are hematoxylin and eosin stained sections illustrating the different tumor regions from the primary ovarian mass as well as disseminated disease in the peritoneum from which genomic DNA was selectively extracted for analysis. **a** Patient a is a 16-year-old female who underwent resection of synchronous bilateral ovarian immature teratomas and debulking of disseminated peritoneal disease. **b** Patient e is a 29-year-old female who underwent resection of a unilateral ovarian immature teratoma and debulking of disseminated peritoneal disease. **c** Patient g is a 25-year-old female who underwent resection of a unilateral ovarian immature teratoma and then four years later underwent resection of a contralateral ovarian mature teratoma (no immature component present).

Fig. 2 Ovarian immature teratomas are characterized by extensive genomic loss of heterozygosity. Plots of ∆allele frequency (∆AF) were generated from the whole exome sequencing data for each of the 52 tumor regions from 10 patients with ovarian immature teratomas. Two distinct tumor clones were identified in four patients (b, g, h, and j) who all had bilateral ovarian teratomas. While all tumor regions harbored a 2N diploid or near-diploid genome, extensive genomic loss of heterozygosity was observed in each of the different tumor components analyzed. Each point represents one informative polymorphic locus. Points near the top of the y-axis represent single nucleotide polymorphisms that are homozygous in the tumor, whereas points near the bottom of the y-axis are heterozygous. y-axis, ∆AF. x-axis, chromosome. Dotted line, centromere. ∆AF is calculated as the absolute difference between theoretical heterozygosity (AF=0.5).

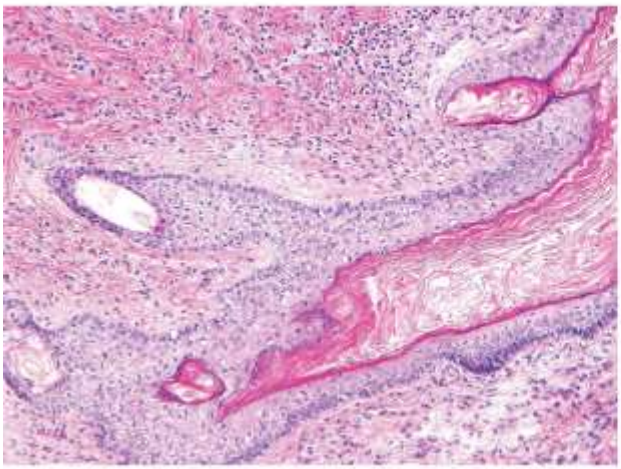
Fig. 3 Identical patterns of genomic loss of heterozygosity among all mature, immature, and disseminated components in ovarian teratomas confirm a single clonal origin, except in females with bilateral tumors. Plots of ∆allele frequency (∆AF) were generated from the whole exome sequencing data for each of the 7 different tumor regions from patient h, an 8-year-old girl who initially underwent resection of a 17 cm immature teratoma from the left ovary, and then 9 years later underwent resection of a 16 cm immature teratoma

from the right ovary as well as debulking of disseminated disease in the peritoneum (gliomatosis peritonei). While all tumor regions harbored a diploid genome, extensive genomic loss of heterozygosity was observed in each of the different tumor components. The immature teratoma and two mature teratoma regions studied from the left ovary had the identical pattern of allelic imbalance, whereas the immature teratoma and two mature teratoma regions studied from the right ovary shared an identical pattern of allelic imbalance that was distinct from the tumor elements in the contralateral ovary. Additionally, the gliomatosis peritonei had an identical pattern of allelic imbalance as the immature and mature teratoma components from the right ovary. Each point represents one informative polymorphic locus. Points near the top of the y-axis represent single nucleotide polymorphisms that are homozygous in the tumor, whereas points near the bottom of the y-axis are heterozygous. y-axis, ΔAF. x-axis, chromosome. Dotted line, centromere. ΔAF is calculated as the absolute difference between theoretical heterozygosity (AF=0.5).

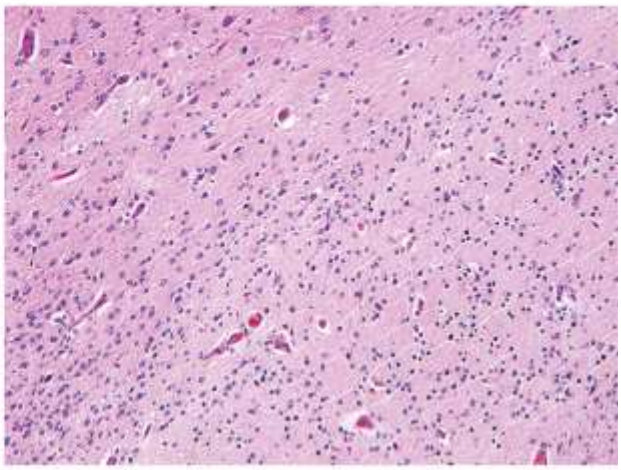
Fig. 4 The five proposed genetic mechanisms of origin of ovarian teratomas from a germ cell. One homologous chromosome pair undergoing two genetic crossing over events is illustrated for simplicity. Orange arrows depict aberrant outcomes of meiosis. Black arrows depict the normal path through meiosis. Each plot depicts a simulated example of the chromosomal loss-of-heterozygosity pattern that arises from each of the five hypothetical mechanisms of origin, measured by the allele frequency difference of SNPs in the tumor compared to constitutional DNA. Y-axis: the two possible zygosity states in a diploid cell (top = homozygosity, bottom = heterozygosity). X-axis: position along an individual chromosome. Vertical dotted blue line depicts the centromere. For Mechanism V, the homozygosity pattern on each chromosome will vary based on the number and location of crossing over events. Adapted from Surti et al. [25].

Patient a

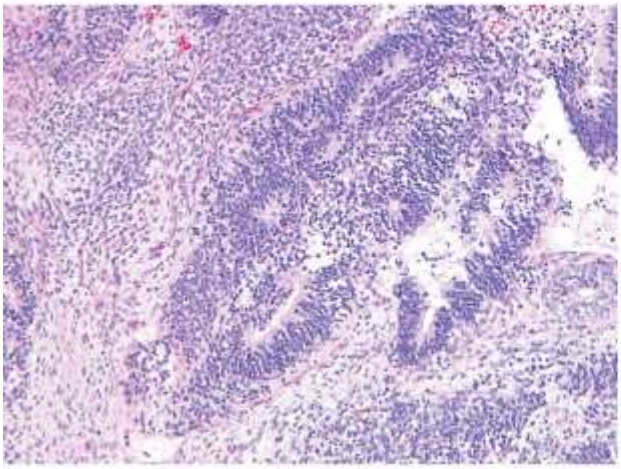
Fig. 1A. Heskett et al. 2019



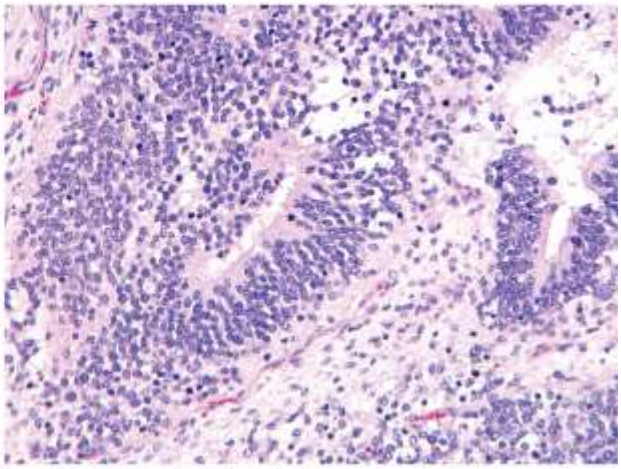
L ovary: mature teratoma, epidermal differentiation



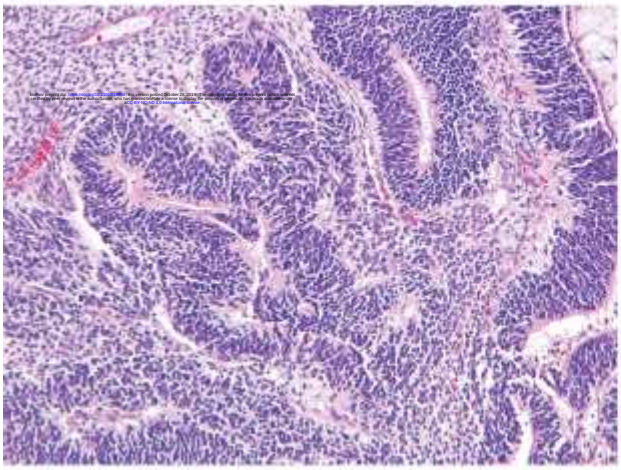
L ovary: mature teratoma, neuroglial differentiation



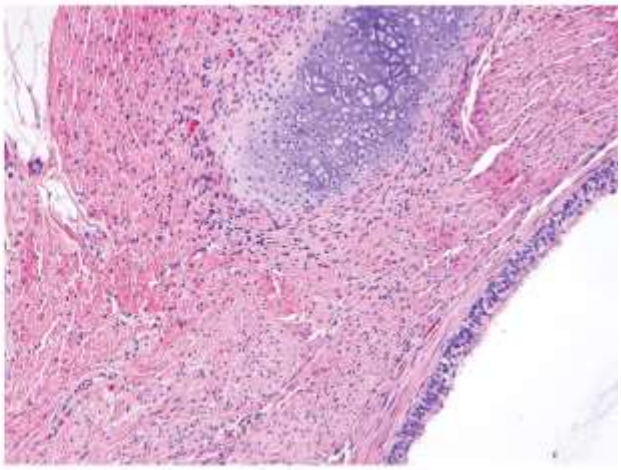
L ovary: immature teratoma



L ovary: immature teratoma



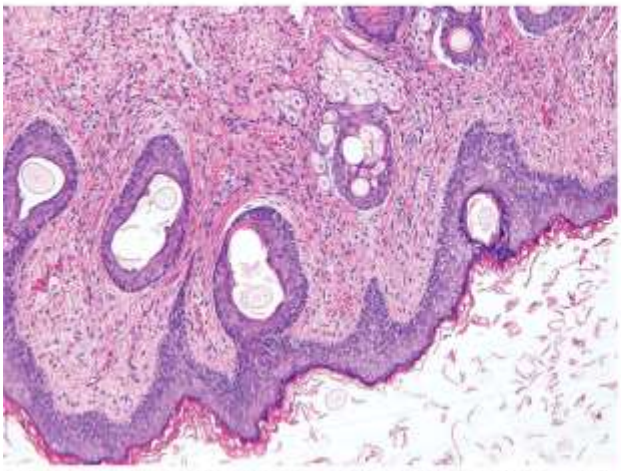
Peritoneum: disseminated immature teratoma



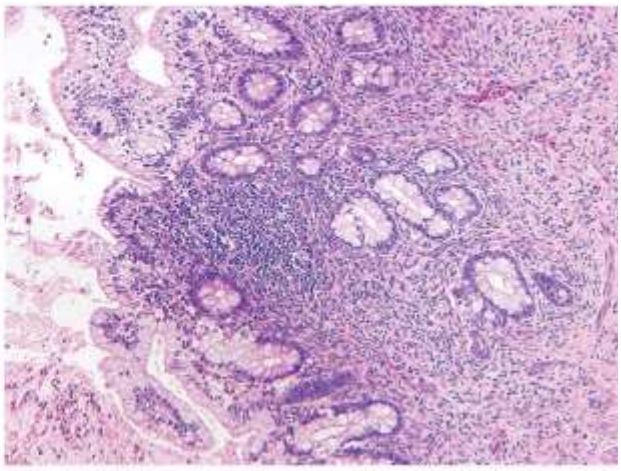
Peritoneum: disseminated mature teratoma

Patient e

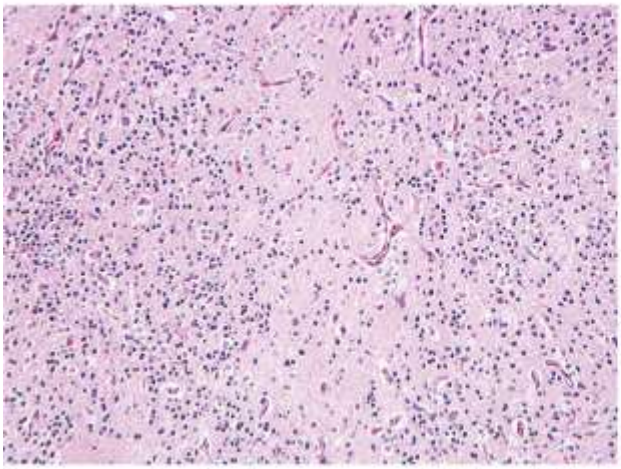
Fig. 1b. Heskett et al 2019



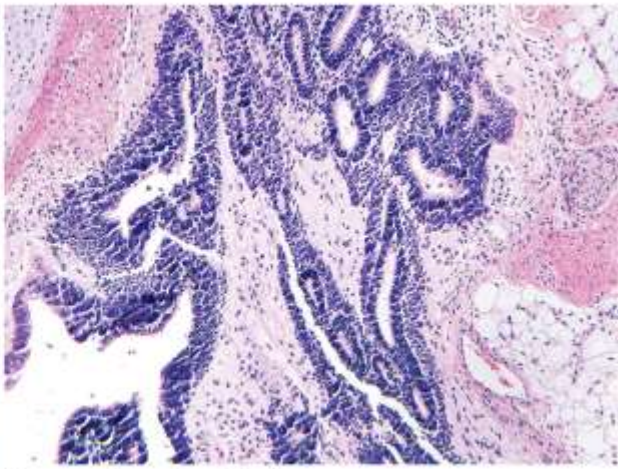
R ovary: mature teratoma, epidermal differentiation



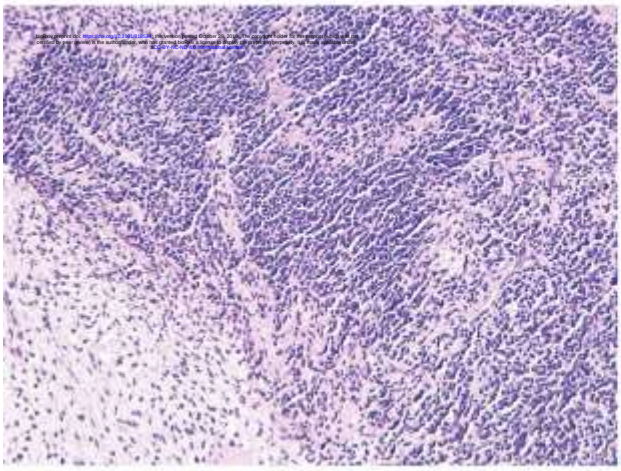
R ovary: mature teratoma, colonic differentiation



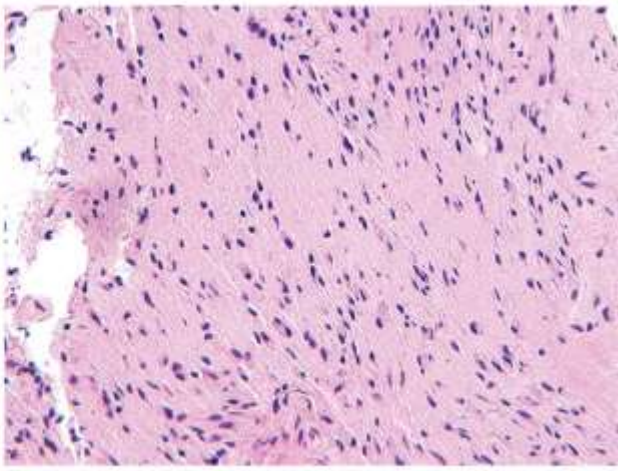
R ovary: mature teratoma, neuroglial differentiation



R ovary: immature teratoma



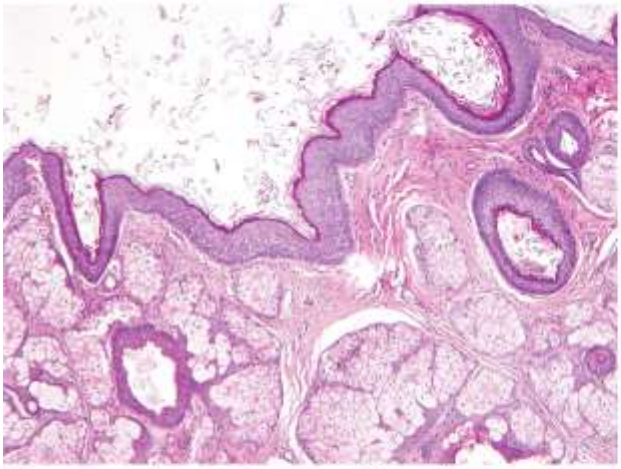
Peritoneum: disseminated immature teratoma



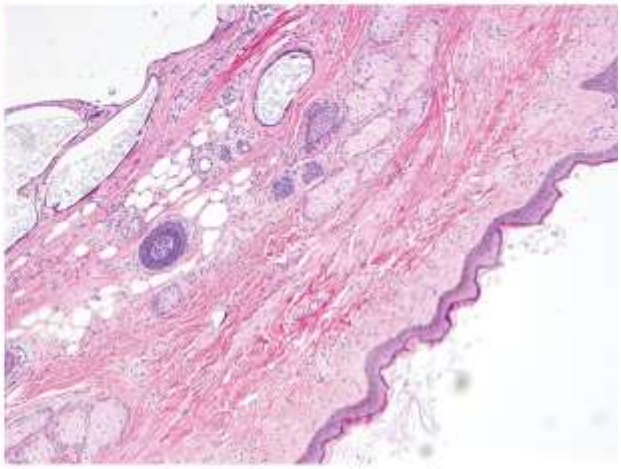
Peritoneum: disseminated mature teratoma (gliomatosis peritonei)

Patient g

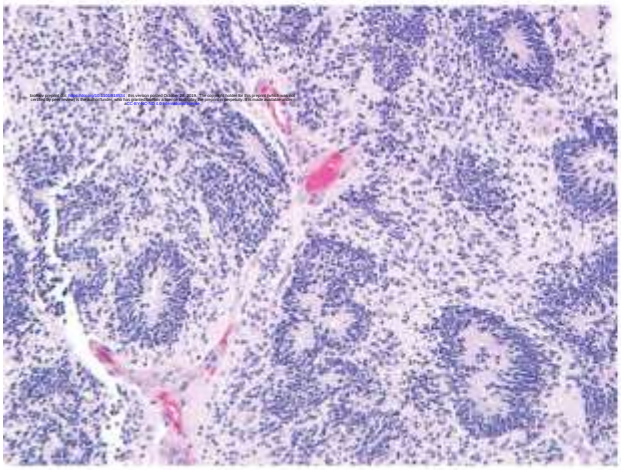
Fig. 1C. Heskett et al. 2019



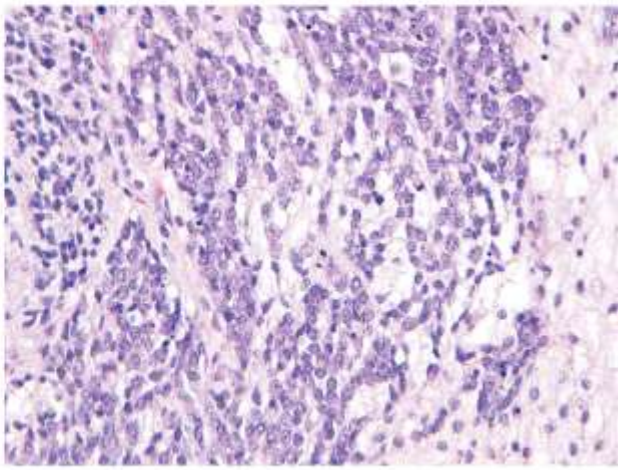
R ovary: mature teratoma, epidermal differentiation



L ovary: mature teratoma, epidermal differentiation



L ovary: immature teratoma



L ovary: minute foci of yolk sac tumor

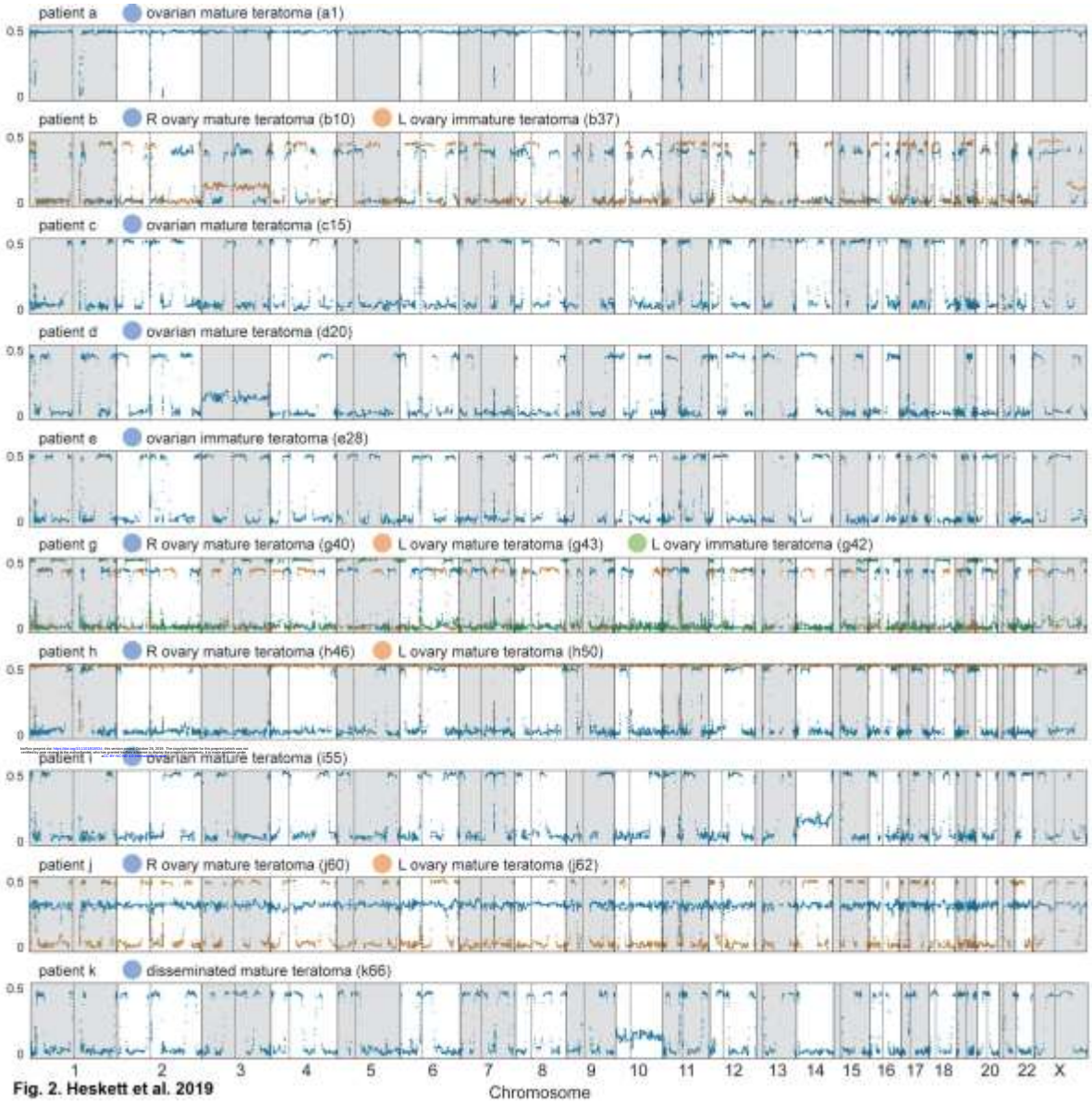


Fig. 2. Heskett et al. 2019

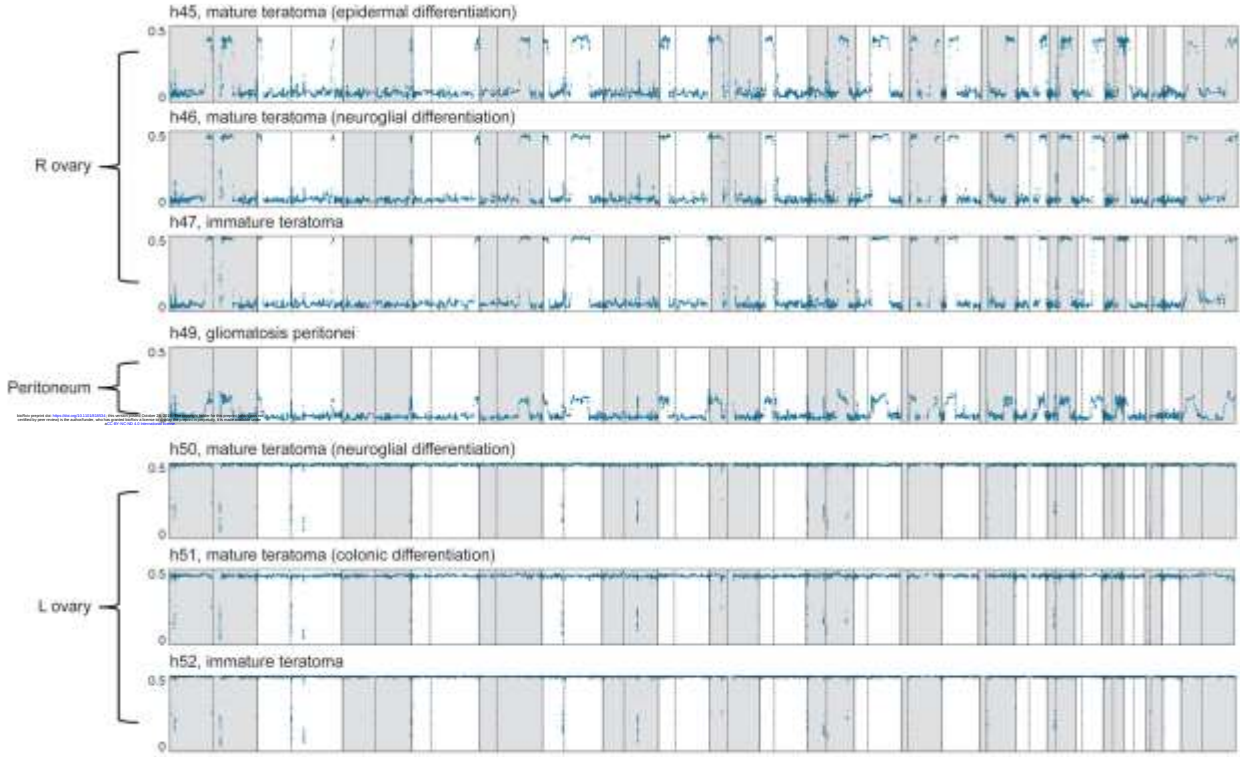


Fig. 3. Heskett et al. 2019

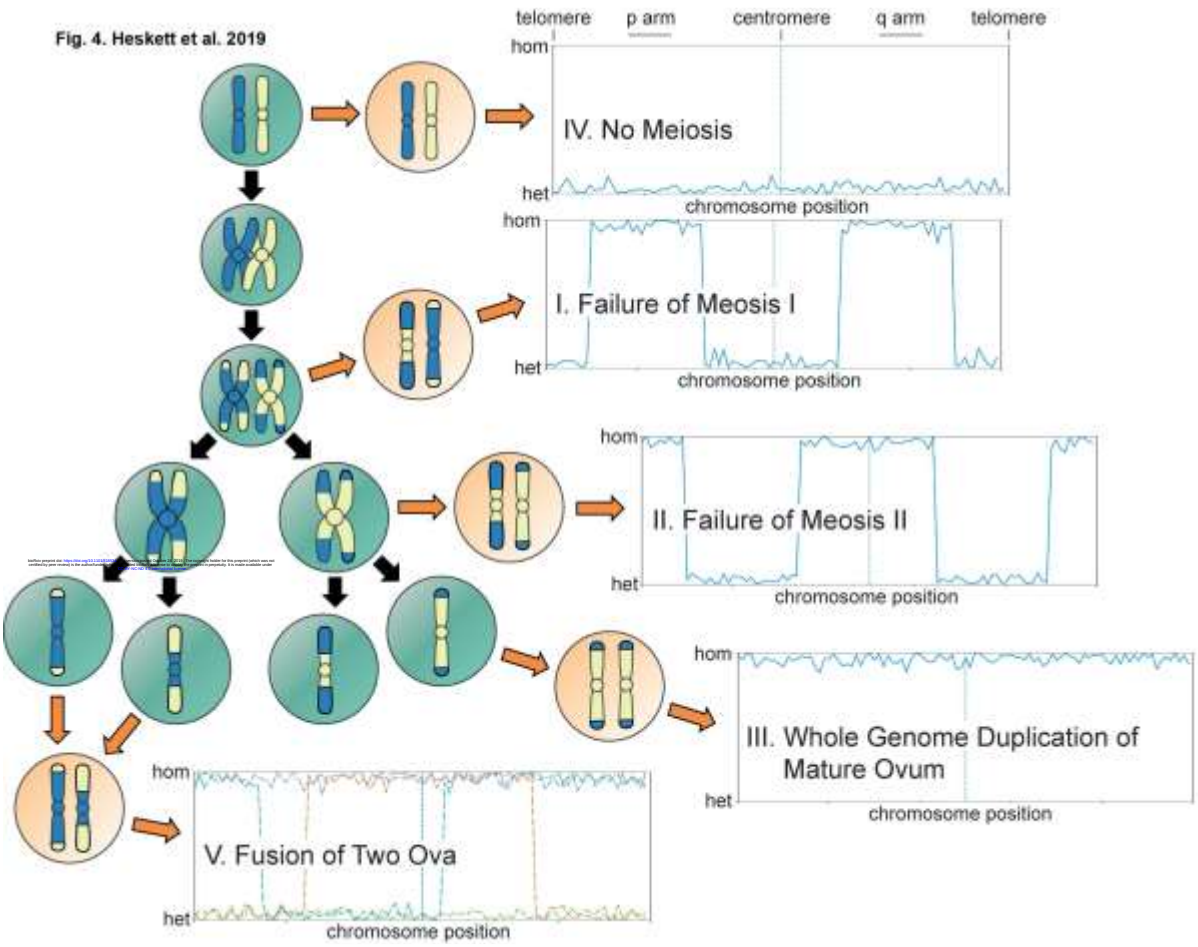


Fig. 4. Heskett et al. 2019
telomere
p arm
centromere
q arm
telomere
hom
het
IV. No Meiosis
chromosome position
I. Failure of Meosis I
II. Failure of Meosis II
III. Whole Genome Duplication of Mature Ovum
V. Fusion of Two Ova

2: Identification and classification of allele-restricted long non-coding RNA that are required for normal chromosome function

2.1 Abstract

Although most of the 3 billion base pair human genome is consistently transcribed into RNA, only 1-2% codes for protein, leaving the structure and function of the vast majority of the genome not readily interpretable. Most of the non-coding transcriptome is classified as non-coding RNA, and while not as well studied as the protein coding RNAs, have been recently found to affect the cellular phenotype through a variety of mechanisms in *cis* and *trans* including chromatin regulation, direct DNA interactions, regulation of coding RNA transcription and mRNA transcripts, as well as acting in nuclear scaffolds and condensates. Two novel, functional long non-coding RNAs, ASAR6 and ASAR15, are non-coding genes of special interest due to the remarkable observation that they are required for essential chromosome behavior including proper chromosome condensation and DNA replication timing on 6 and 15. Here, we test the hypothesis that all autosomes contain ASAR genes responsible for normal DNA replication timing, and utilize the special characteristics of ASARs, including nuclear restricted expression of a monoallelic transcript, to find additional ASAR genes. By leveraging allele-specific transcriptional and replication timing sequencing data, we report a set of potential ASAR genes across all autosomes, perform gold-standard validation of expression, and show that allele-specific deletion of ASAR genes results in aberrant DNA replication timing. Evidence is provided supporting a model whereby ASAR long non-coding RNA are expressed across autosomes and regulate normal DNA replication.

2.2 Introduction

Composition of the noncoding genome

As early as 1971[33] it was estimated that only 6% of the 3 billion base pairs of human DNA content was utilized for protein coding genes. More recent large scale sequencing and annotation efforts have revealed that approximately 1-2% of the human genome codes for protein while roughly 80% is actively transcribed and regulated [34], posing a major challenge in the field genome biology to describe the form and function of the billions of base pairs of transcribed sequence. The majority of the non-protein coding transcriptome is considered to be noncoding RNA (ncRNA), a diverse group that includes the well-known and highly conserved ribosomal RNA(rRNA) and transfer RNA (tRNA), as well as several other classifications of ncRNA of varying length from tens of nucleotides (nt) to many thousands. ncRNAs less than 200nt in length are defined as short ncRNAs and include the microRNAs (miRNAs) that achieve post-transcriptional regulation of mRNA, small nucleolar RNAs(snoRNAs) that guide post-transcriptional modification of rRNAs and tRNAs, and piwiRNAs (piRNAs) known for binding with piwi proteins to suppress transposable element activity in germline cells, among many others. The description and function of small ncRNAs have been the subject of numerous recent reviews [35], but the focus of this discussion is on the role of long ncRNAs.

Long noncoding RNAs

Canonical long non-coding RNAs (lncRNAs) are defined as expressed RNAs greater than 200nt with no known protein coding potential. The bulk of known annotated lncRNAs share similarities with coding transcripts including undergoing splicing, similar size of mature transcript (~1.3kb), presence of 5'-caps and 3'-poly A tails, presence of epigenetic marks H3K4me3 at active promoters and H3K36me3 in gene bodies, and transcription by RNA polymerase II (RNAPII).

Across the genome, lncRNAs may overlap coding genes in an antisense orientation, exist entirely within an intron antisense to a coding gene, be adjacent to a coding gene but diverge in direction, or be wholly intergenic to protein coding regions (Figure 2.1). An additional line of emerging research shows that a class of previously undiscovered very long intergenic noncoding RNAs (vlincRNAs) may comprise ~10% of the genome and are abundant in disease and normal cell models and primary human tissues [36]. vlincRNAs are defined as being at least 50kb in length of continuously transcribed RNA that are not associated with protein coding genes and contain a high concentration of transposable elements.

Mechanisms of lncRNA function

The progress towards elucidation of ncRNA function has suffered from several inherent disadvantages compared to protein coding genes. lncRNA are less conserved across species and display regions of conservation surrounded by unconstrained sequence making predictions of functional significance more difficult. lncRNA are also expressed at significantly lower levels than coding mRNAs, requiring selective RNA-sequencing (RNA-seq) library preparation and computational methods. Furthermore, 3-dimensional secondary structure predictions are hindered by the long length and dynamic conformations that RNA molecules may occupy[37].For example, the critical A-repeat region of Xist, the well-studied lncRNA on the X-chromosome that achieves dosage compensation in female mammalian cells, is believed to bind to multiple proteins that facilitate transcriptional repression of the X-chromosome, has been extensively studied but remains unsolved with low agreement between multiple models[37]. Early high throughput attempts to infer the functions of lncRNAs found distinct lncRNA expression profiles in different spatial regions of the mouse brain, consistent with the idea that lncRNAs were not merely

transcriptional noise or an artifact of chromatin remodeling as previously proposed, but instead have function as active RNA elements [38]. More recent studies have utilized a tool kit of direct and indirect methods to predict lncRNA functional importance including differential expression studies, "guilt by association" inference, condition-specific expression, disease association, evolutionary conservation, cellular localization, epigenetic status, lncRNA locus and lncRNA interactions with protein, and lncRNA/ nucleic acid binding (reviewed in: [39]. Figure 2.2).

In fact, lncRNAs are able to affect the expression of protein coding genes through a variety of mechanisms in *cis* or *trans* including interacting with chromatin complexes[40], modulating proteins and enzyme cofactors[41], recruiting DNA/RNA binding proteins [42], competitive antisense transcription, forming a triple helix with DNA[43], and participating in high-order structures[44-46] (Figure 2.3). In *cis,* modulation of neighboring gene expression can occur as a result of actions dependent on the lncRNA product itself, or the act of transcription leading to reduced recruitment of RNAPII to protein coding gene promoters[46]. In *trans,* lncRNAs can interact with chromatin modifying complexes capable of changing the epigenetic landscape of chromosomes, interact with protein and enzyme cofactors to change the activity of proteins affecting diverse cellular processes, and regulate gene expression transcriptionally and post-transcriptionally by binding to DNA-binding proteins and RNA-binding proteins, respectively. Interestingly, recent evidence has shown that 3-dimensional genome structure in the nucleus may be regulated by lncRNAs acting as scaffolds that could control the dynamics of liquid-phase separated components. On the X chromosome in mammalian females, the lncRNA Xist orchestrates a complex restructuring and compaction of nearly the entire chromosome by recruitment of the polycomb repressive complex and other RNA binding proteins resulting in the strong repression of ~1,000 genes[45].

Notable examples of lncRNA with known functions

Many lncRNAs have been linked to monoallelic gene expression. For example, the 2.7kb H19 RNA on chromosome 11 is monoallelically expressed from the maternal allele at a location adjacent to the paternally expressed IGF2 gene, a major driver of cell growth. H19 and IGF2 are reciprocally expressed from opposite homologs with IGF2 being expressed from the paternal chromosome, and this relationship is required for normal embryonal and placental development [47]. H19 has also been implicated in multiple cancer types where it is involved in cancer gene regulation via miRNA sponge effects and additional yet unknown mechanisms resulting from aberrant increased or decreased expression (Wang, oncol letters 2020). As mentioned above, lncRNA may be expressed at complex loci forming distinct expression patterns with linked genes. Kcnq1 is a coding gene expressed from a complex imprinted domain where the paternally expressed antisense lncRNA Kcnq1ot1 is expressed from within an intron of the Kcnq1 protein coding gene, silencing the paternal copy and several neighboring genes during early embryogenesis [48]. Likewise, disruption of Kcnq1ot1 leads to improper embryonal heart development in mice [48]. The Air (or AIRN) noncoding RNA on chromosome 6 in humans is also regulated by parent of origin specific imprinting. Predominantly expressed from the paternal allele Air expression leads to embryonic silencing of the IGF2R gene *in cis*, as well as silencing of two other genes Slc22a2 and Slc22a3 in the placenta[49]. Air has been found to recruit the H3K9 histone methyltransferase to the promoter of Slc22a3 in the placenta, silencing transcription by repressive histone-modifying activities[49]. MALAT1 (or NEAT2) is an >8kb lncRNA on chromosome 11 that is highly conserved and expressed among 33 mammalian species[50]. MALAT1 is abundant in nuclear speckles and interacts with proteins involved in splicing[51].

MALAT1 as a promoter or repressor of metastasis has been observed with differing results dependent on cancer type[52].

The X-inactivation system of chromosome remodeling and gene repression is orchestrated by Xist

One of the first discovered and arguably the most well studied lncRNA is Xist, the centerpiece of X chromosome inactivation (XCI). XCI establishes dosage compensation in mammals and represents a chromosome-wide system of allele-specific gene expression [53]. In early female embryos X chromosome inactivation occurs as epiblast cells differentiate into the embryonic germ layers during gastrulation, and this choice is maintained as the parent cell of different lineages divides and goes on to form tissues throughout the body, leading to approximately 50% ratio of inactivation of each X homolog in mature tissues [54]. Most of the knowledge generated about XCI has been generated in the mouse model. Mouse embryonic stem cells (mESC) contain two active X chromosomes and will undergo random XCI when the cells are differentiated *in vitro [53]*. XCI demonstrates *sensing*, where XCI is only triggered if a cell contains at least two X chromosomes, and *choice*, the ability to select between inactivation of one homolog or the other. The genetic locus known as the X-inactivation center (Xic) is necessary for XCI, and harbors the Xist gene which gets spliced, retained in the nucleus, and is capable of coating the chromosome from which it is expressed creating a uniquely organized heterochromatic environment [55]. Xist RNA contains at least 5 conserved tandem repeat domains that are conserved between human and mouse, required for proper XCI, which serve to recruit effector proteins[45].

Although the exact mechanism of Xist silencing is still under investigation, it is known that Xist works in tandem with RNA binding proteins such as polycomb repressive complex (PRC) and heterogeneous nuclear ribonucleoprotein U (hnRNPU) [53](Figure 2.4). SMRT and SPEN are

recruited to the inactive X and activate histone deacetylases, resulting in deactylated histone proteins and more tightly bound DNA [45]. RNA binding proteins including hnRNPU, UCHL5, EXOSC5, and RBM15 are known to bind to the distinct domains of Xist and are involved in stability, processing, and silencing[56]. After XCI is triggered on one homolog, promoter methylation of Xist on the opposite homolog is present and suggests an epigenetic repression mechanism. Adjacent to Xist are several other long non-coding RNAs that have roles in XCI. Tsix[57], a long noncoding RNA located 13kb downstream of Xist and organized antisense, is a negative regulator of Xist expression and is associated with the active X chromosome, demonstrating an example of a nonfunctional antisense gene that is critical to the function of a nearby functional gene[54]. Tsix is theorized to act by several plausible mechanisms: formation of duplex RNA with Xist and subsequent degradation, competition for RNA binding proteins that are necessary for Xist function, or modification of Xist chromatin after Tsix antisense expression. The lncRNA Jpx, encoded 10kb upstream and antisense to Xist, has coordinated expression with Xist, and is proposed to act *in trans* or *in* cis to activate Xist.

Repetitive elements comprise a large fraction of the human genome and are present across mammalian evolution

Upon the initial sequencing of the human genome it was reported that ~55% of the genome is comprised of repetitive DNA sequences, with more recent estimates increasing to two thirds [58, 59]. The majority of repetitive DNA sequences belong to the broad family of transposable elements including retrotransposons that are capable of propagating throughout the genome by a "copy and paste" mechanism. Retrotransposons contain two main groups, long terminal repeat (LTR) elements and non-LTR elements. Non-LTR elements include the long and short interspersed

nuclear elements (LINEs and SINEs). Full length LINEs are 6-8kb and contain two open reading frames that encode proteins sufficient for autonomous retrotransposition. Active LINE-1 elements (L1s) have a life cycle whereby two open reading frames (ORF1 and ORF2) are transcribed and the RNA is exported to the cytoplasm. Next, the ORF1 and ORF2 are translated to proteins and will bind L1 RNA to form a ribonucleoprotein complex that is imported into the nucleus. At the genomic target site, L1 endonuclease will nick the target DNA, and L1 retrotranscriptase will reverse transcribe from the L1 RNA to create DNA that is integrated into the genome[60]. LINE-1 elements are ubiquitous across eukaryotes and highly active in mammalian species, making up ~18% of the human genome [61]. In mammals, L1 elements participate in an "evolutionary arms race" [62] by which active L1s are repressed by the host, and then evolve to bypass the repression, leading to a linear pattern of evolution whereby only the youngest L1s are active. In non-mammalian vertebrates, L1s make up far less of the content genome (~0.5%), however multiple divergent L1 element types are active in a given genome[34].

L1s and other transposable elements have several hypothesized roles in genome evolution. Although transposable elements appear to be selfish and have no direct benefit on the host genome, they can affect gene expression by inserting into an exon, intron, or regulatory region, and may catalyze a variety of genomic rearrangements of including deletions, duplications, translocations and inversions [63]. Transposable element mobility is not considered a major driver of disease but there are many notable exceptions in cancer where ORF1p and ORF2p are overexpressed highly in tumor types such as breast, ovarian, and pancreatic[64]. Disruptive L1s transposition into a tumor suppressor gene is a plausible mechanism of retrotransposon mediated tumorigenesis, though not commonly observed, have been found in the APC gene in colon cancer patients [64].

LINE1 elements are enriched on the X chromosome and play a role in X-inactivation

Another unique characteristic of the X chromosome is the high fraction of sequence derived from L1 retrotransposons compared to all autosomes [65]. L1 elements comprise about 17% of the human genome and contain a 5' and 3' untranslated region and two open reading frames that encode for ORF1p and ORF2p, both required for retrotransposition[66]. L1s do not contain long terminal repeats and mobilize through an RNA dependent copy and paste mechanism. Most L1 elements in the human genome are non-functional owing to disruptive 5' truncations, inversions, or point mutations within the L1 encoded open reading frames[66], and the rate of active transposition events in somatic and germline tissues is under investigation. One recent study found that evolutionarily young LINEs located within introns of protein coding genes serve to recruit RNA-binding proteins to repress RNA splicing at cryptic sites, whereas evolutionarily older LINEs are found closer to exons and serve to enhance RNA processing in a tissue specific manner[67]. These findings illuminate a potential mechanism for the evolution of RNA processing and may explain the phenomenon of aberrant RNA splicing in disease.

X has twice as many L1 elements compared to autosomes, with the highest density clustering located at the X inactivation center (XIC), and lowest density at regions that escape X inactivation. It has been proposed that L1 elements act as "boosters" of X-inactivation [68], and although this hypothesis remains unproven, one line of evidence using RNA/DNA FISH visualization has shown participation of silent evolutionarily older LINEs in assembly of the heterochromatin environment, and younger active LINEs mediating local propagation of XCI into regions that are adjacent to escape regions.

ASARs are vlincRNAs that are required for normal DNA replication and chromosome stability

A family of two vlincRNA were recently discovered by performing a genetic screen to find rearranged chromosomes that display dysfunctional DNA replication timing and mitotic condensation phenotypes, an unexplained phenomenon that has long been observed in tumor derived genomes.

ASAR6 and ASAR15, located on chromosome 6 and 15 retrospectively, encode for ~200kb non-canonical lncRNAs that are not spliced, not polyadenylated, and transcribed by RNAPII. ASAR6 is intergenic and closely linked to here coding genes, MANEA, FUT9, and FHL5 [69, 70]. ASAR15 resides at a complex locus on 15q24 that also encodes a protein coding gene SCAPER and microRNA3713. Both ASAR vlincRNAs display differential allelic expression that can be visualized with fluorescent in-situ hybridization (FISH) where one large signal is observed, retained within the nucleus and within the territory of the chromosome from which they are expressed[71-73](Figure 2.5).

A notable feature of ASAR6 and ASAR15 is that they display programmed random monoallelic expression (PRME). RME is a form of allele specific expression that is defined by a random choice of allelic expression that may be followed by stable mitotic transmission. Perhaps the most well studied example of RME is on the X chromosome of female mammalian cells, where PRME of most genes on the chromosome solves the issue of dosage imbalance between cells of male and female origin. On autosomes, PRME is only well understood for specialized classes of genes in the nervous and immune system. RME of ASARs was established by DNA/RNA FISH and RT-PCR followed by sequencing at heterozygous SNPs in peripheral blood lymphocytes in unrelated individuals[69, 71]. It is currently unknown when and how monoallelic expression is established for ASARs. Evidence shows that the ASAR6 locus is required for mono-allelic expression of the linked protein coding genes FUT9 and FHL5 in the fibrosarcoma cell line HTD114 [70], and

follow up experiments are warranted to determine the significance of the ASAR6/MANEA/FUT9/FHL5 monoallelic expression/asynchronous replication unit. In addition to their unusual localization pattern and mode of expression, ASARs appear to be critical for normal chromosome function. Deletion or disruption of either of two lincRNA genes ASAR6 or ASAR15, respectively, leads to a delay in mitotic chromosome condensation (DMC) of whole chromosomes as well as a delay in replication timing (DRT) which ultimately leads to chromosome instability, a perilous phenotype that may result in decreased integrity of the genome. While the information about the mechanism of ASAR function is currently limited, there is [74] evidence that the ASARs mediate their effects in an RNA dependent manner. Ectopic integration of ASAR6 onto mouse chromosome 3 leads to the DRT/DMC phenotype, and this phenotype is subsequently rescued by antisense oligonucleotides (locked nucleic acid gapmers) that knock down the RNA [74]. Interestingly, it was found that the genetic deletion of a single antisense, full length L1PA2 transposable element within ASAR6 in HTD114 cells led to a substantial DRT phenotype, suggesting a critical function of the L1PA2-containing RNA product in replication timing. Intriguingly, loss of function of ASARs on human chromosomes and gain of function (ectopic integration) on mouse chromosomes leads to the same phenotype, demonstrating a dominant negative action. Because ASARs are not only expressed from the early allele, ASARs are not simply effectors of early replication. Aberrant DNA replication timing has been observed

The discovery of ASAR genes in a tissue culture model system gives some hint to their function and potential significance. ASAR genes were originally found during the investigation of a long unexplained phenotype in cancer cells of abnormally condensed chromosomes during mitosis. To search for the cause of DMC, rearranged chromosomes displaying the DMC phenotype were studied [75]. A series of microcell hybrid panels containing chromosomes from the

rhabdomyosarcoma cell line RH30 and the small-cell lung carcinoma cell line CRL-5485 were generated and revealed that multiple rearranged chromosomes with 3q translocations exhibited DMC. DMC was validated epigenetically by the observation of delay of histone H3 phosphorylation during mitosis, a canonical marker of chromosome condensation[75]. Chromosome condensation during mitosis is necessary for the movement of chromosomes on the mitotic spindle and segregation of sister chromatids, and is thought to resolve the problem of chromosome tangling, while creating a substantial chromosome size reduction that allows for efficient movement to the daughter cell before the end of mitosis[76]. Thus, a chromosome undergoing DMC would be at risk of segregation errors, known as nondisjunction, if mitotic condensation was not corrected leading to cells with altered genomes and instability[77].

Genomic methods to study DNA replication timing, lncRNA expression and function, and chromosome structure

Basic molecular methods to study replication timing have traditionally used PCR or DNA-FISH methods that utilize temporal sampling to distinguish between early, late, and asynchronous replication. The single-dot, double-dot assay distinguishes between synchronous and asynchronous replication by counting the ratio of cells in S-phase with one versus two hybridization signals. One signal indicates one allele has replicated and the other has not, whereas two signals indicate both alleles have replicated, thus a high ratio of one-dot cells would indicate replication asynchrony. Replication Timing-Specific Hybridization (RETISH) [69] utilizes a FISH hybridization probe that only binds to DNA that has incorporated BRDU, and has been used to compare a small number of asynchronous loci. RETISH gives a quantitative measurement of asynchrony but doesn't inform the temporal duration of asynchrony, and is laborious enough to

prevent scalable use. These and similar methods are considered gold-standard for individual loci but fail in their ease of scalability to more than one genomic locus at a time and offer limited information about the absolute temporal position of replication of the analyzed locus.

Repli-seq [78] is a next generation sequencing method that utilizes BRDU incorporation in cycling cells followed by FACS sorting based on DNA content, anti-BRDU immunoprecipitation to enrich newly replicated DNA, and short read sequencing to asses replication timing on a genome-wide level. Repli-seq has been performed on two sorted fractions (early vs late), 6 fractions (late G1, S1, S2, S3, S4, early G2), on 16 fractions to improve resolution (S1-S16), and on single cells where the protocol simply relies on measuring excess reads in each cell in lieu of BRDU incorporation and immunoprecipitation.

The identification and annotation of the entire non-coding transcriptome including lncRNA has posed a unique challenge owing to issues including low expression of lncRNAs compared to protein coding mRNA leading to an underrepresented sampling, poor understanding of sequence to function relationships, and lesser conservation of lncRNAs compared to protein coding genes. Collectively these issues prevent accurate transcriptomic measurements without special methods designed to enrich sequencing libraries, and greatly reduce accuracy of predictive models based on presence of functional sequences and high evolutionary conservation that have succeeded in the coding genome.

Search for ASARs across all autosomes reveals autosomal stability centers

ASAR6 [70] and ASAR15 [71] were found by examining the chromosomal regions surrounding translocation breakpoints on a small number of engineered rearranged chromosomes that exhibit DRT/DMC, and there is strong speculation that other similar genes exist. First, chromosomes from

three sources displayed the ASAR deletion phenotype DRT/DMC: 1) cancer cell lines and primary tumor samples, 2) generated with a Cre/LoxP recombination system, 3) created by ionizing radiation. In the Cre/loxP system, the frequency of rearranged chromosomes that display DRT/DMC was consistently found to be ~5%. Assuming a uniform distribution of possible breakpoints throughout the genome, and that every translocation has two products, this implies that roughly 2.5% of the genome could contain ASAR-like elements necessary for normal replication timing and chromosome stability. One could additionally hypothesize that since the known lncRNAs essential for chromosome behavior are Xist, ASAR6, and ASAR15, other autosomes as well may contain a similar gene that controls replication timing of the whole chromosome. Alternatively, if ASAR genes regulate DNA replication timing domains which exist as <1mb regions throughout the genome, it is plausible that each domain is controlled by an ASAR gene, in which case thousands of ASARs could exist. Since asynchronous replication is not limited to the X chromosome, ASARs may be controllers of monoallelic expression/asynchronous replication domains, however these loci have not been systematically studied and our knowledge is currently limited to a small number of regions with anecdotal evidence in specific embryonic developmental stages, or linked to specialized genetic units such as olfactory or immune receptors. The research presented here will test the hypothesis that additional ASAR genes exist on all autosomes. Significant challenges have prevented the testing of this hypothesis by prior research mainly driven by the lack of an appropriate model system. First there is the challenge of assessing monoallelic or differential allelic expression across the genome which is only solved if a haplotype-resolved reference genome is available. Despite the decrease in cost for deep whole-genome sequencing, haplotype resolution remains a challenge. Statistical inference of haplotypes can be performed utilizing population level reference data that has recently been made available.

Haplotypes can be accurately determined by pedigree-based methods, but require genotyping of multiple members of families, and these samples are rarely available. Lastly, creation of long bacterial artificial chromosome (BAC) inserts can allow direct sequencing of individual haplotypes, but this method is laborious and not scalable to high numbers of samples [79].

Next generation RNA-sequencing methods are extensively used for whole-transcriptome measurements, however because ASAR RNAs are retained in the nucleus, non-poly adenylated, and as with other lncRNA genes more lowly expressed than protein coding transcripts, achieving significant sequencing depth for discerning between alleles adds cost and preprocessing steps. Because selection for poly-adenylated (pA) transcripts is not performed, RNA samples have the potential to be significantly contaminated by ribosomal RNA, furthering the need for more deeply sequenced libraries. Secondly, to discern between random monoallelic expression, genetic influence on expression by expressed quantitative trait loci (eQTLs), dynamic differential expression, and imprinted regions, the analysis of cell populations from multiple developmental linages, or "clones", within one individual is required. Since the definition of a clonal population varies in different contexts and may not present with easily measurable biomarkers, determining the clonality of cell line model systems has remained a challenge. Prior to the research presented here, no known haplotype-resolved, clonally derived model systems were available. Confirmation of random monoallelic expression requires observation of expression from opposite alleles within the same individual, or if only one sample is available per individual, differential allelic expression from different alleles within a population at the single cell level. Lastly, asynchronous replication has only been studied in limited contexts [80, 81]. Current methods to asses asynchronous replication include the single-dot double-dot assay [82] where the ratio of cells in S-phase that

show one versus two FISH signals at a given locus allows a quantitative measurement of a single locus, but this method does not easily scale to more than a small number of loci.

By utilizing clonally derived, haplotype resolved model systems, allele specific repli-seq, and deep RNA-seq, we circumvent previous scientific challenges and present novel findings autosomal asynchronous replication domains that express mono-allelic lncRNAs. We then observe the DRT/DMC phenotype after allele specific deletion of the active lncRNA loci, consistent with previous behavior of ASAR genes. Given these results, we reveal a new feature of autosomes whereby X-chromosome-like domains, possibly regulated by lncRNA, may serve a critical role in stability and replication, warranting further investigation into the mechanism and significance of these loci.

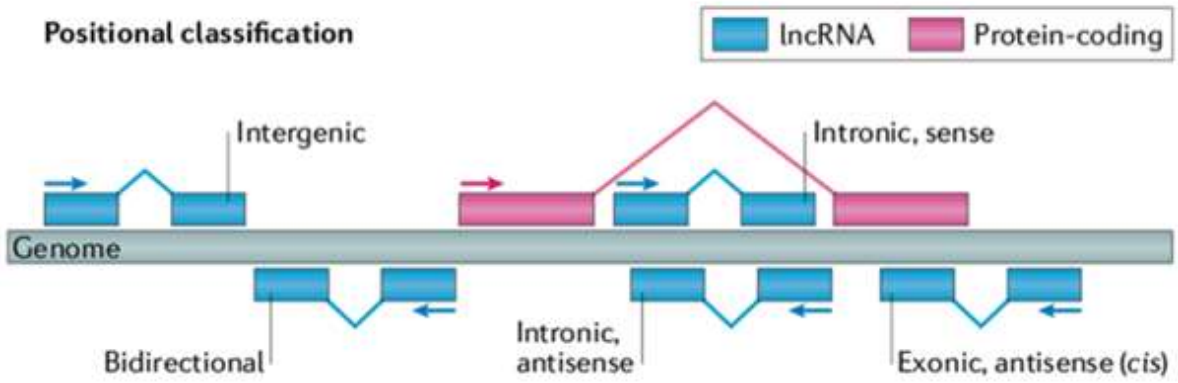


Figure 2.1. The genomic organization of lncRNAs. lncRNAs may exist distant from protein coding genes (intergenic) in a sense or antisense direction, entirely within introns (intronic, sense or antisense), or overlap coding exons in an antisense direction. Figure adapted from [83].

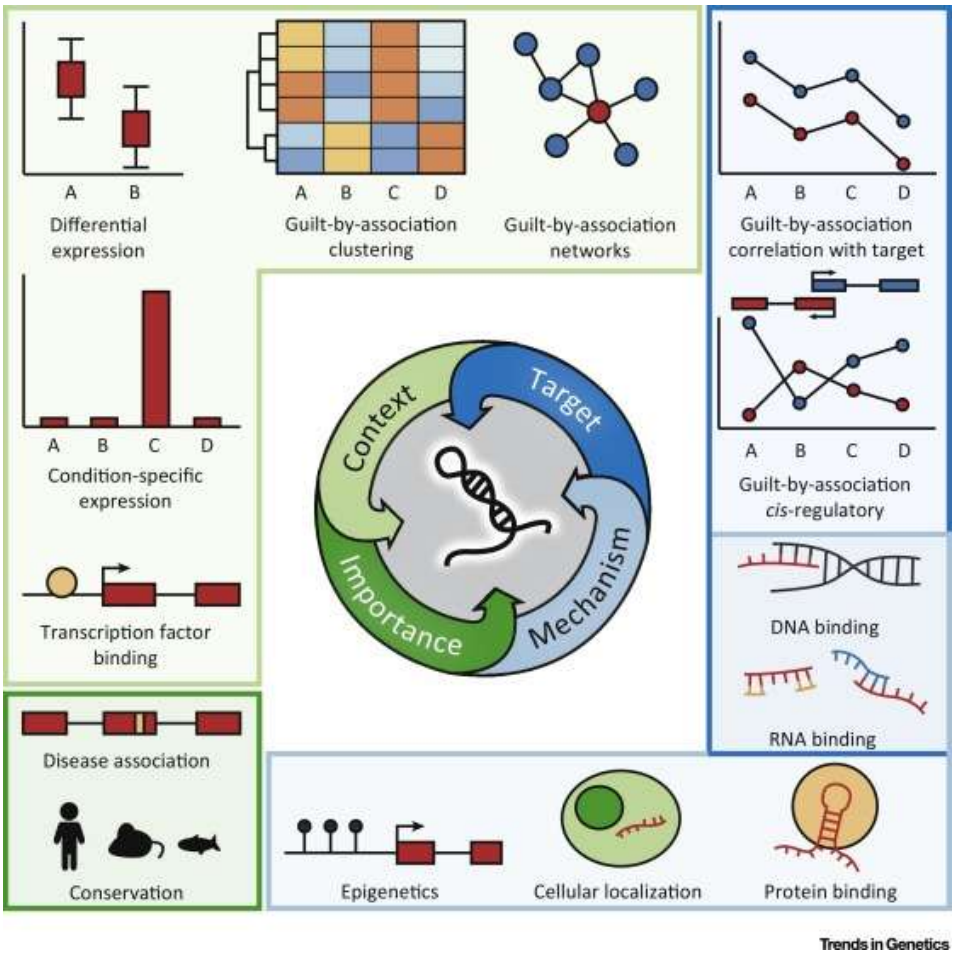


Figure 2.2. Array of methods used to predict and determine lncRNA function and significance. Light green: informative of biological context. Dark green: biological importance across species. Dark blue: informative of potential regulatory context. Light blue: informative of functional mechanism. (From [39])

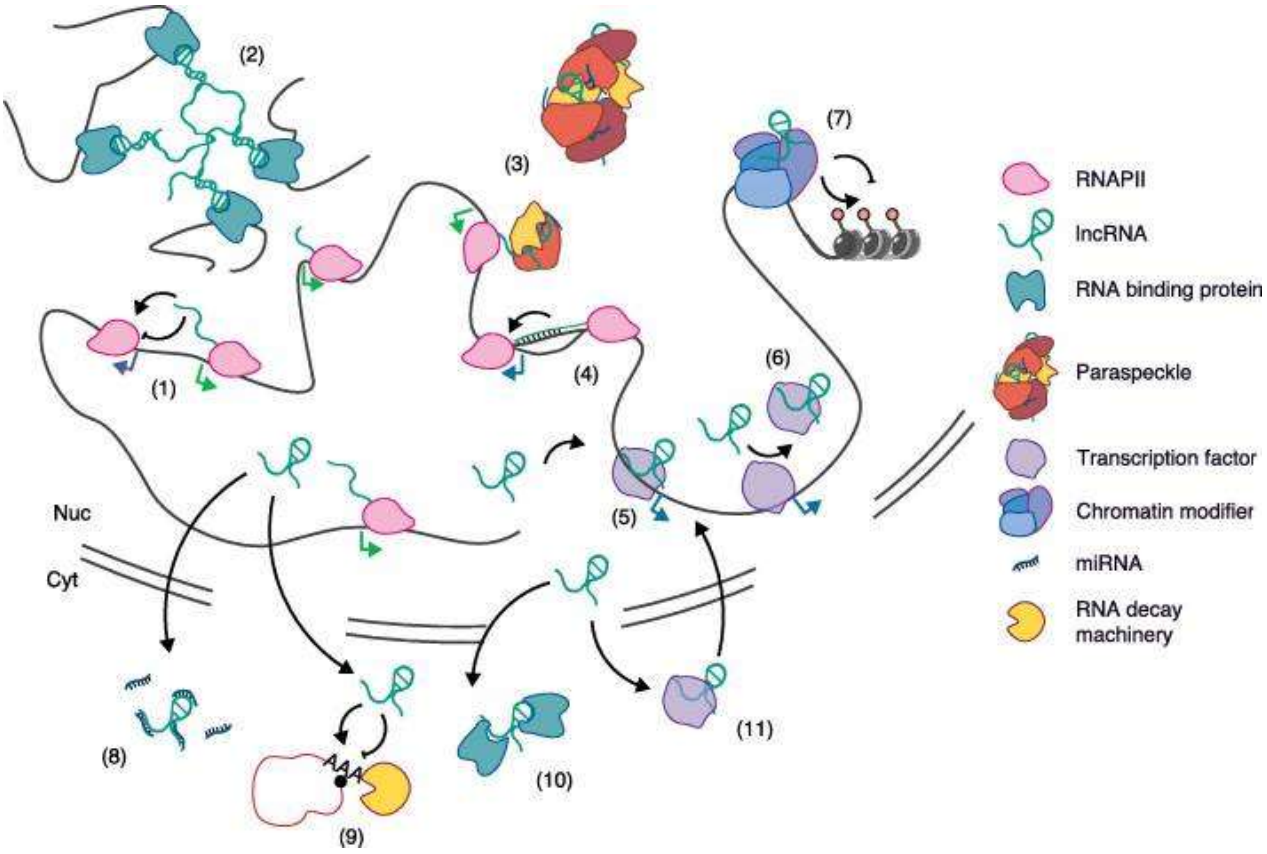


Figure 2.3. The known and suspected mechanisms of lncRNA function. 1) Gene expression of neighboring genes is repressed or activated in a lncRNA transcription dependent manner. 2) lncRNA facilitate interchromosomal interactions. 3) lncRNA as a component of nuclear structures including paraspeckles. 4) DNA/RNA triplex formation. 5) lncRNA binding to DNA binding proteins such as transcription factors as a guide. 6) lncRNA binding to transcription factors in a repressive manner as a decoy. 7) lncRNA as a component of a chromatin remodeling complex. 8) lncRNA binds to microRNA leading to increased microRNA degradation. 9) lncRNA may regulate

stability of mRNA. 10-11) lncRNA binds to DNA and RNA binding proteins which can affect cellular localization. (Figure adapted from [46])

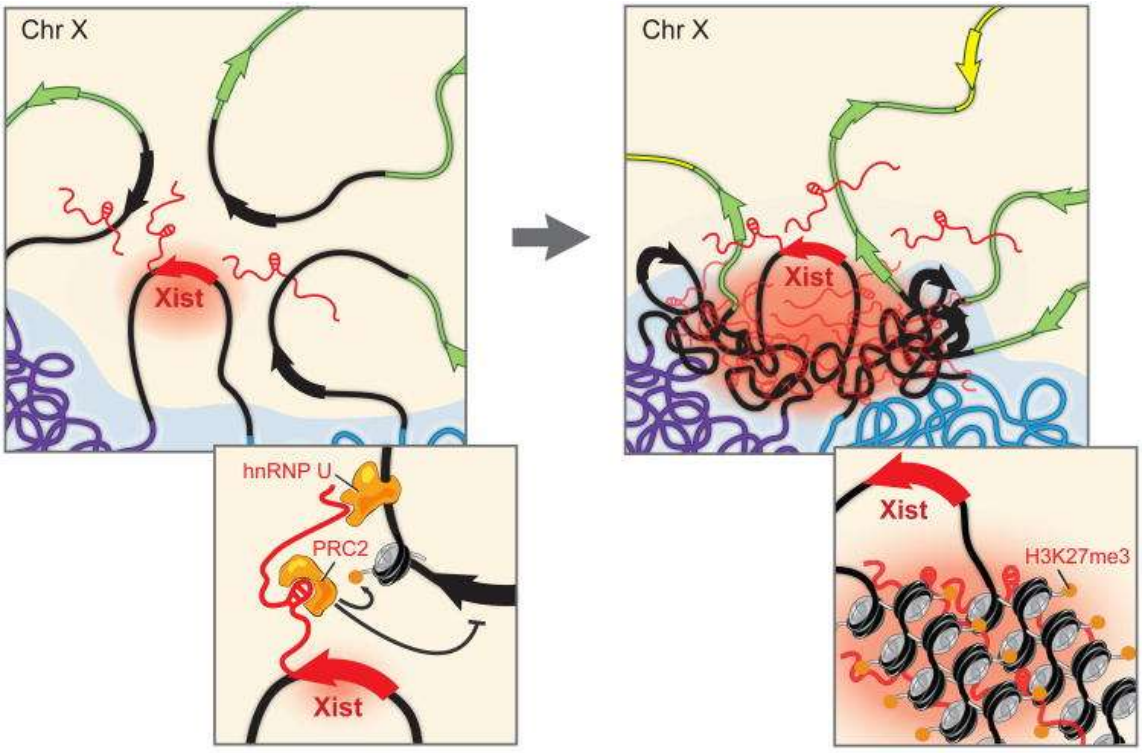


Figure 2.4. Model of Xist mediated X-inactivation inferred from DNA/RNA hybridization approaches. Left: Xist is expressed and interacts with gene-rich DNA from spatially proximal regions, dependent on the A-rich domain of Xist as well as recruitment of chromatin remodeling protein PRC2 and hnRNP U. Right: Gene dense regions are compacted into an Xist compartment that displays enrichment of the inactive mark H3K27me3. Figure adapted from [84].

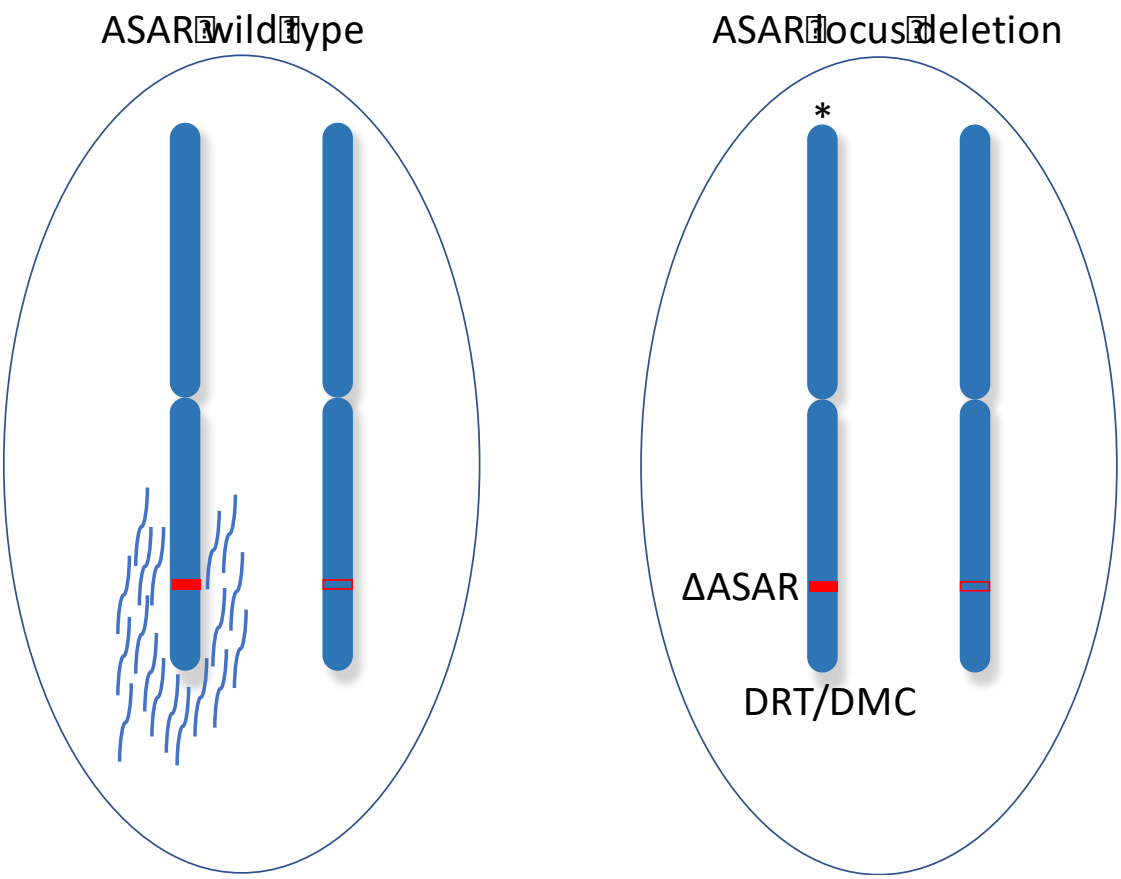


Fig 2.5. Basic model of allele specific ASAR lncRNA expression. Left: ASAR is expressed mono-allelically and produces a vlncRNA that remains in the close territory of the chromosome from which it is expressed (red: ASAR locus). Right: The functional consequence of an allele specific deletion of the ASAR locus on the active allele results in the DRT/DMC phenotype and an unstable chromosome.

Chapter 2: Works Cited

1. Scultetus, J., *Trichiasis admiranda sive Morbus pilaris mirabilis observatus.* 1658.
2. Damjanov, I., B.B. Knowles, and D. Solter, *The Human Teratomas: Experimental and Clinical Biology*. 2012: Humana Press.
3. Kurman, R.J., *WHO classification of tumours of female reproductive organs.* 2014.
4. Shen, H., et al., *Integrated Molecular Characterization of Testicular Germ Cell Tumors.* Cell Rep, 2018. **23**(11): p. 3392-3406.
5. Williamson, S.R., et al., *The World Health Organization 2016 classification of testicular germ cell tumours: a review and update from the International Society of Urological Pathology Testis Consultation Panel.* Histopathology, 2017. **70**(3): p. 335-346.
6. Richardson, B.E. and R. Lehmann, *Mechanisms guiding primordial germ cell migration: strategies from different organisms.* Nat Rev Mol Cell Biol, 2010. **11**(1): p. 37-49.
7. Gilbert, S.F., *Developmental Biology, 6th Edition*. 2000.
8. Smith, H.O., et al., *Incidence and survival rates for female malignant germ cell tumors.* Obstet Gynecol, 2006. **107**(5): p. 1075-85.
9. Møller, H. and H. Evans, *Epidemiology of gonadal germ cell cancer in males and females.* Apmis, 2003. **111**(1): p. 43-6; discussion 46-8.
10. Kraggerud, S.M., et al., *Molecular characteristics of malignant ovarian germ cell tumors and comparison with testicular counterparts: implications for pathogenesis.* Endocr Rev, 2013. **34**(3): p. 339-76.

11. Dasari, S. and P.B. Tchounwou, *Cisplatin in cancer therapy: molecular mechanisms of action.* Eur J Pharmacol, 2014. **740**: p. 364-78.

12. Solheim, O., et al., *Malignant ovarian germ cell tumors: presentation, survival and second cancer in a population based Norwegian cohort (1953-2009).* Gynecol Oncol, 2013. **131**(2): p. 330-5.

13. Hersmus, R., et al., *The biology of germ cell tumors in disorders of sex development.* Clin Genet, 2017. **91**(2): p. 292-301.

14. Litchfield, K., et al., *Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor.* Nat Genet, 2017. **49**(7): p. 1133-1140.

15. Norris, H.J., H.J. Zirkin, and W.L. Benson, *Immature (malignant) teratoma of the ovary: a clinical and pathologic study of 58 cases.* Cancer, 1976. **37**(5): p. 2359-72.

16. Alwazzan, A.B., et al., *Pure Immature Teratoma of the Ovary in Adults: Thirty-Year Experience of a Single Tertiary Care Center.* Int J Gynecol Cancer, 2015. **25**(9): p. 1616-22.

17. Liang, L., et al., *Gliomatosis peritonei: a clinicopathologic and immunohistochemical study of 21 cases.* Mod Pathol, 2015. **28**(12): p. 1613-20.

18. Snir, O.L., et al., *Frequent homozygosity in both mature and immature ovarian teratomas: a shared genetic basis of tumorigenesis.* Mod Pathol, 2017. **30**(10): p. 1467-1475.

19. Hoei-Hansen, C.E., et al., *Ovarian dysgerminomas are characterised by frequent KIT mutations and abundant expression of pluripotency markers.* Mol Cancer, 2007. **6**: p. 12.

20. Palmer, R.D., et al., *Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets.* Cancer Res, 2010. **70**(7): p. 2911-23.

21. Linder, D., *Gene loss in human teratomas.* Proceedings of the National Academy of Sciences of the United States of America, 1969. **63**(3): p. 699-704.

22. Surti, U., et al., *Genetics and biology of human ovarian teratomas. I. Cytogenetic analysis and mechanism of origin.* American journal of human genetics, 1990. **47**(4): p. 635-643.

23. Parrington, J.M., L.F. West, and S. Povey, *The origin of ovarian teratomas.* Journal of medical genetics, 1984. **21**(1): p. 4-12.

24. Alberts, B.J.A.L., J., *Molecular Biology of the Cell. 4th edition.* 2002.

25. Haenel, Q., et al., *Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics.* Mol Ecol, 2018. **27**(11): p. 2477-2497.

26. McGranahan, N. and C. Swanton, *Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.* Cell, 2017. **168**(4): p. 613-628.

27. Durinck, S., et al., *Temporal dissection of tumorigenesis in primary cancers.* Cancer Discov, 2011. **1**(2): p. 137-43.

28. Gerstung, M., et al., *The evolutionary history of 2,658 cancers.* Nature, 2020. **578**(7793): p. 122-128.

29. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-892.

30. Quirk, J.T., N. Natarajan, and C.J. Mettlin, *Age-specific ovarian cancer incidence rate patterns in the United States.* Gynecol Oncol, 2005. **99**(1): p. 248-50.

31. Shaaban, A.M., et al., *Ovarian malignant germ cell tumors: cellular classification and clinical and imaging features.* Radiographics, 2014. **34**(3): p. 777-801.

32. Carlson, B.M., *Gametogenesis.* Reference Module in Biomedical Sciences, 2014.

33. Kimura, M., *Theoretical foundation of population genetics at the molecular level.* Theoretical Population Biology, 1971. **2**(2): p. 174-208.

34. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

35. Gebert, L.F.R. and I.J. MacRae, *Regulation of microRNA function in animals.* Nature Reviews Molecular Cell Biology, 2019. **20**(1): p. 21-37.

36. St Laurent, G., et al., *VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer.* Genome Biology, 2013. **14**(7): p. R73.

37. Jones, A.N. and M. Sattler, *Challenges and perspectives for structural biology of lncRNAs—the example of the Xist lncRNA A-repeats.* Journal of Molecular Cell Biology, 2019. **11**(10): p. 845-859.

38. Mercer, T.R., et al., *Specific expression of long noncoding RNAs in the mouse brain.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 716-21.

39. Signal, B., B.S. Gloss, and M.E. Dinger, *Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs.* Trends in Genetics, 2016. **32**(10): p. 620-637.

40. Grote, P., et al., *The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse.* Dev Cell, 2013. **24**(2): p. 206-14.

41. Liu, B., et al., *A cytoplasmic NF-kappaB interacting long noncoding RNA blocks IkappaB phosphorylation and suppresses breast cancer metastasis.* Cancer Cell, 2015. **27**(3): p. 370-81.

42. Huarte, M., et al., *A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response.* Cell, 2010. **142**(3): p. 409-19.

43. Boque-Sastre, R., et al., *Head-to-head antisense transcription and R-loop formation promotes transcriptional activation.* Proc Natl Acad Sci U S A, 2015. **112**(18): p. 5785-90.

44. Lewandowski, J.P., et al., *The Firre locus produces a trans-acting RNA molecule that functions in hematopoiesis.* Nature Communications, 2019. **10**(1): p. 5137.

45. Loda, A. and E. Heard, *Xist RNA in action: Past, present, and future.* PLoS Genet, 2019. **15**(9): p. e1008333.

46. Marchese, F.P., I. Raimondi, and M. Huarte, *The multidimensional mechanisms of long noncoding RNA function.* Genome Biology, 2017. **18**(1): p. 206.

47. Zhang, Y. and B. Tycko, *Monoallelic expression of the human H19 gene.* Nature Genetics, 1992. **1**(1): p. 40-44.

48. Korostowski, L., N. Sedlak, and N. Engel, *The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart.* PLoS Genet, 2012. **8**(9): p. e1002956.

49. Nagano, T., et al.

50. Amodio, N., et al., *MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches.* Journal of hematology & oncology, 2018. **11**(1): p. 63-63.

51. Arun, G., D. Aggarwal, and D.L. Spector, *MALAT1 Long Non-Coding RNA: Functional Implications.* Non-coding RNA, 2020. **6**(2): p. 22.

52. Sun, Y. and L. Ma, *New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis.* Cancers, 2019. **11**(2): p. 216.

53. Augui, S., E.P. Nora, and E. Heard, *Regulation of X-chromosome inactivation by the X-inactivation centre.* Nature Reviews Genetics, 2011. **12**(6): p. 429-442.

54. Plath, K., et al., *Xist RNA and the Mechanism of X Chromosome Inactivation.* Annual Review of Genetics, 2002. **36**(1): p. 233-278.

55. Clemson, C.M., et al., *XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure.* J Cell Biol, 1996. **132**(3): p. 259-75.

56. Weidmann, C.A., et al., *Analysis of RNA–protein networks with RNP-MaP defines functional hubs on RNA.* Nature Biotechnology, 2021. **39**(3): p. 347-356.

57. Lee, J.T., L.S. Davidow, and D. Warshawsky, *Tsix, a gene antisense to Xist at the X-inactivation centre.* Nat Genet, 1999. **21**(4): p. 400-4.

58. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

59. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

60. Viollet, S., C. Monot, and G. Cristofari, *L1 retrotransposition: The snap-velcro model and its consequences.* Mobile genetic elements, 2014. **4**(1): p. e28907-e28907.

61. Ivancevic, A.M., et al., *LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life.* Genome Biology and Evolution, 2016. **8**(11): p. 3301-3322.

62. Boissinot, S. and A. Sookdeo, *The Evolution of LINE-1 in Vertebrates.* Genome biology and evolution, 2016. **8**(12): p. 3485-3507.

63. Warren, I.A., et al., *Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates.* Chromosome Res, 2015. **23**(3): p. 505-31.

64. Burns, K.H., *Transposable elements in cancer.* Nature Reviews Cancer, 2017. **17**(7): p. 415-424.

65. Bailey, J.A., et al., *Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(12): p. 6634-6639.

66. Beck, C.R., et al., *LINE-1 elements in structural variation and disease.* Annual review of genomics and human genetics, 2011. **12**: p. 187-215.

67. Attig, J., et al., *Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing.* Cell, 2018. **174**(5): p. 1067-1081.e17.

68. Lyon, M.F., *X-Chromosome inactivation: a repeat hypothesis.* Cytogenetic and Genome Research, 1998. **80**(1-4): p. 133-137.

69. Donley, N., et al., *Asynchronous replication, mono-allelic expression, and long range Cis-effects of ASAR6.* PLoS Genet, 2013. **9**(4): p. e1003423.

70. Stoffregen, E.P., et al., *An autosomal locus that controls chromosome-wide replication timing and mono-allelic expression.* Hum Mol Genet, 2011. **20**(12): p. 2366-78.

71. Donley, N., L. Smith, and M.J. Thayer, *ASAR15, A cis-acting locus that controls chromosome-wide replication timing and stability of human chromosome 15.* PLoS Genet, 2015. **11**(1): p. e1004923.

72. Donley, N. and M.J. Thayer, *DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability.* Semin Cancer Biol, 2013. **23**(2): p. 80-9.

73. Thayer, M.J., *Mammalian chromosomes contain cis-acting elements that control replication timing, mitotic condensation, and stability of entire chromosomes.* BioEssays, 2012. **34**(9): p. 760-770.

74. Platt, E.J., L. Smith, and M.J. Thayer, *L1 retrotransposon antisense RNA within ASAR lncRNAs controls chromosome-wide replication timing.* J Cell Biol, 2018. **217**(2): p. 541-553.

75. Smith, L., A. Plug, and M. Thayer, *Delayed replication timing leads to delayed mitotic chromosome condensation and chromosomal instability of chromosome translocations.* Proc Natl Acad Sci U S A, 2001. **98**(23): p. 13300-5.

76. Koshland, D. and A. Strunnikov, *MITOTIC CHROMOSOME CONDENSATION.* Annual Review of Cell and Developmental Biology, 1996. **12**(1): p. 305-333.

77. Chang, B.H., et al., *Chromosomes with delayed replication timing lead to checkpoint activation, delayed recruitment of Aurora B and chromosome instability.* Oncogene, 2007. **26**(13): p. 1852-61.

78. Hansen, R.S., et al., *Sequencing newly replicated DNA reveals widespread plasticity in human replication timing.* Proceedings of the National Academy of Sciences, 2010. **107**(1): p. 139.

79. Snyder, M.W., et al., *Haplotype-resolved genome sequencing: experimental methods and applications.* Nature Reviews Genetics, 2015. **16**(6): p. 344-358.

80. Bartholdy, B., et al., *Allele-specific analysis of DNA replication origins in mammalian cells.* Nature communications, 2015. **6**: p. 7051-7051.

81. Rivera-Mulia, J.C., et al., *Allele-specific control of replication timing and genome organization during development.* Genome Res, 2018. **28**(6): p. 800-811.

82. Ensminger, A.W. and A. Chess, *Coordinated replication timing of monoallelically expressed genes along human autosomes.* Human Molecular Genetics, 2004. **13**(6): p. 651-658.

83. Uszczynska-Ratajczak, B., et al., *Towards a complete map of the human long non-coding RNA transcriptome.* Nat Rev Genet, 2018. **19**(9): p. 535-548.

84. Engreitz, J.M., et al., *The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.* Science (New York, N.Y.), 2013. **341**(6147): p. 1237973-1237973.

2.3: Reciprocal Monoallelic Expression of ASAR lncRNA genes controls replication timing of human chromosome 6. Heskett. Et al. 2020.

Reciprocal monoallelic expression of ASAR lncRNA genes controls replication timing of human chromosome 6.

Michael Heskett[1], Leslie G. Smith[2], Paul Spellman[1], and Mathew J. Thayer[2]*

[1]Department of Chemical Physiology and Biochemistry
Oregon Health & Science University
3181 S. W. Sam Jackson Park Road
Portland, Oregon 97239, USA

[2]Department of Chemical Physiology and Biochemistry
Oregon Health & Science University
3181 S. W. Sam Jackson Park Road
Portland, Oregon 97239, USA

***Corresponding author.**
ORCID: 0000-0001-6483-1661
TEL: (503) 494-2447
FAX: (503) 494-7368
Email:thayerm@ohsu.edu

Running Title: ASAR lincRNA genes on chromosome 6

Key Words: Replication timing, cis-acting element, non-coding RNA

Abstract

DNA replication occurs on mammalian chromosomes in a cell-type distinctive temporal order known as the replication timing program. We previously found that disruption of the noncanonical lncRNA genes *ASAR6* and *ASAR15* results in delayed replication timing and delayed mitotic chromosome condensation of human chromosome 6 and 15, respectively. *ASAR6* and *ASAR15* display random monoallelic expression, and display asynchronous replication between alleles that is coordinated with other random monoallelic genes on their respective chromosomes. Disruption of the expressed allele, but not the silent allele, of *ASAR6* leads to delayed replication, activation of the previously silent alleles of linked monoallelic genes, and structural instability of human chromosome 6. In this report, we describe a second lncRNA gene (*ASAR6-141*) on human chromosome 6 that when disrupted results in delayed replication timing in *cis*. *ASAR6-141* is subject to random monoallelic expression and asynchronous replication, and is expressed from the opposite chromosome 6 homolog as *ASAR6*. ASAR6-141 RNA, like ASAR6 and ASAR15 RNAs, contains a high L1 content and remains associated with the chromosome territory where it is transcribed. Three classes of *cis*-acting elements control proper chromosome function in mammals: origins of replication, centromeres; and telomeres, which are responsible for replication, segregation and stability of all chromosomes. Our work supports a fourth type of essential chromosomal element, "Inactivation/Stability Centers", which express ASAR lncRNAs responsible for proper replication timing, monoallelic expression, and structural stability of each chromosome.

Author summary

Mammalian cells replicate their chromosomes during a highly ordered and cell type-specific program. Genetic studies have identified two long non-coding RNA genes, *ASAR6* and *ASAR15*, as critical regulators of the replication timing program of human chromosomes 6 and 15, respectively. There are several unusual characteristics of the ASAR6 and ASAR15 RNAs that distinguish them from other long non-coding RNAs, including: being very long (>200 kb), lacking splicing of the transcripts, lacking polyadenylation, and being retained in the nucleus on the chromosomes where they are made. *ASAR6* and *ASAR15* also have the unusual property of being expressed from only one copy of the two genes located on homologous chromosome pairs. Using these unusual characteristics shared between *ASAR6* and *ASAR15*, we have identified a second ASAR lncRNA gene located on human chromosome 6, which we have named *ASAR6-141*. *ASAR6-141* is expressed from the opposite chromosome 6 homolog as *ASAR6*, and disruption of the expressed allele results in delayed replication of chromosome 6. ASAR6-141 RNA had previously been annotated as vlinc273. The very long intergenic non-coding (vlinc)RNAs represent a recently annotated class of RNAs that are long (>50 kb), non-spliced, and non-polyadenlyated nuclear RNAs. There are currently >2,700 vlincRNAs expressed from every chromosome, are encoded by >15% of the human genome, and with a few exceptions have no known function. Our results suggest the intriguing possibility that the vlinc class of RNAs may be functioning to control the replication timing program of all human chromosomes.

Introduction

Numerous reports over the past 50+ years have described an abnormal DNA replication phenotype affecting individual chromosomes in mitotic preparations from mammalian cells [1]. For example, we found that certain tumor derived chromosome translocations display a delay in replication timing (DRT) that is characterized by a >3 hour delay in the initiation and completion of DNA synthesis along the entire length of individual chromosomes [2]. Chromosomes with DRT also display a delay in mitotic chromosome condensation (DMC), which is characterized by an under-condensed appearance during mitosis and a concomitant delay in the mitotic phosphorylation of histone H3 [2, 3]. We have also found that ~5% of chromosomal translocations induced by exposing human cells to ionizing radiation (IR) display DRT/DMC [4]. To characterize the DRT/DMC phenotype further, we developed a Cre/loxP system that allowed us to create chromosome translocations in a precise and controllable manner [4, 5]. Using this Cre/loxP system, we carried out a screen in human cells designed to identify loxP integration sites that generate chromosome translocations with DRT/DMC [4-7]. We found that ~5% of Cre/loxP induced translocations display DRT/DMC [5]. Therefore, ~5% of translocations induced by two different mechanisms (IR or Cre/loxP) result in DRT/DMC.

Our Cre/loxP screen identified five cell lines that generate balanced translocations, affecting eight different autosomes, all displaying DRT/DMC [5]. Characterization of two of these translocations identified discrete *cis*-acting loci that when disrupted result in DRT/DMC on human chromosomes 6 or 15 [6, 7]. Molecular examination of the disrupted loci identified two lncRNA genes, which we named ASynchronous replication and

Autosomal RNA on chromosome 6 (*ASAR6*) and on chromosome 15 (*ASAR15*) [6, 7]. These studies defined the first *cis*-acting loci that control replication timing, monoallelic gene expression, and structural stability of individual human autosomes [6, 7].

The vast majority of genes on mammalian autosomes are expressed from both alleles. However, some autosomal genes are expressed preferentially from only one allele, achieving a state of "autosome pair non-equivalence" [8, 9]. The most extreme form of differential allelic expression is often referred to as monoallelic expression, where a single allele is expressed exclusively (reviewed in [10]). Differential allelic expression can arise from distinct mechanisms. For example, differential expression can arise due to DNA sequence polymorphisms within promoter or enhancer elements that influence the efficiency with which a gene will be transcribed (reviewed in [11, 12]). In contrast, differential expression can occur in the absence of DNA sequence polymorphisms and is connected to situations where there is a "programmed" requirement to regulate gene dosage or to provide exquisite specificity (reviewed in [11, 13-15]). One well established form of programmed monoallelic expression occurs in a parent of origin specific manner, and is known as genomic imprinting (reviewed in [16]). In addition, monoallelic expression occurring in a random manner has been observed from as many as 8% of autosomal genes [12, 17]. One unusual characteristic of all programmed monoallelic genes is asynchronous replication between alleles [8, 9, 18, 19]. This asynchronous replication is present in tissues where the genes are not transcribed, indicating that asynchrony is not dependent on transcription [7-9, 20]. Furthermore, asynchronous replication of random monoallelic genes is coordinated with other random monoallelic genes on the same chromosome, indicating that there is a chromosome-wide system that coordinates

replication asynchrony of random monoallelic genes [7-9, 20]. We use the following criteria to classify genes as being subject to Programed Random Monoallelic Expression (PRME): 1) monoallelic expression is detected in multiple unrelated individuals, which rules out rare DNA polymorphisms in promoters or enhancers; 2) monoallelic expression of either allele is detected in single cell-derived subclones from the same individual, which rules out genomic imprinting; and 3) asynchronous replication is present and coordinated with other random monoallelic genes on the same chromosome, indicating that the monoallelic gene is regulated by a chromosome-wide system that coordinates asynchronous replication along chromosome pairs. Using these criteria, we previously found that *ASAR6* and *ASAR15* are subject to PRME [6, 7, 20].

Recent reports have described very long intergenic non-coding (vlinc)RNAs expressed in numerous human tissues [21-23]. The vlincRNAs are RNA Pol II products that are nuclear, non-spliced, non-polyadenylated transcripts of >50 kb of contiguously expressed sequence that are not associated with protein coding genes. The initial reports annotated 2,147 human vlincRNAs from 833 samples in the FANTOM5 dataset [23, 24]. A more recent study identified an additional 574 vlincRNAs expressed in childhood acute lymphoblastic leukemia [25]. Therefore, there are currently >2,700 annotated vlincRNAs that are encoded by >15% of the human genome [23-25]. ASAR6 and ASAR15 RNAs share several characteristics with the vlincRNAs, including: RNA Pol II products, long contiguous transcripts (>50 kb) that are non-spliced, non-polyadenylated, and are retained in the nucleus [6, 7, 20]. Therefore, given these shared characteristics between ASAR6, ASAR15 and vlincRNAs, we consider the vlincRNAs as potential ASAR candidates.

ASAR6 and ASAR15 RNAs also share additional characteristics, including: PRME, are retained within the chromosome territories where they are transcribed, and contain a high long interspersed element 1 (LINE1 or L1) content [6, 7, 20]. In this report, we used these "ASAR" characteristics to identify a second lncRNA gene that controls replication timing of human chromosome 6, which we designate as *ASAR6-141*. The *ASAR6-141* gene is located at ~141 mb of human chromosome 6, is subject to random monoallelic expression and asynchronous replication, and disruption of the expressed allele, but not the silent allele, leads to delayed replication of human chromosome 6 in *cis*. ASAR6-141 RNA, which was previously annotated as vlinc273 [23], is ~185 kb in length, contains ~30% L1 sequences, remains associated with the chromosome 6 territory where it is transcribed, and is expressed in *trans* to the expressed allele of *ASAR6*. These observations support a model that includes reciprocal monoallelic expression of different ASAR lincRNA genes that control replication timing of homologous chromosome pairs [26].

Results

Reciprocal random monoallelic expression of ASAR lncRNAs on chromosome 6: With the goal of identifying nuclear RNAs with "ASAR" characteristics expressed from human chromosome 6, we carried out RNA-seq on nuclear RNA isolated from HTD114 cells. HTD114 cells are a human fibrosarcoma cell line, where we previously carried out the Cre/loxP screen that led to the identification and functional characterization of *ASAR6* and *ASAR15* [5-7]. Figure 1a shows the UCSC Genome Browser view of chromosome 6, between 140.2 mb and 141.3 mb, showing the RNA-seq reads from the region previously annotated as expressing 6 different vlincRNAs [23]. We note that vlinc273 is expressed, but vlinc271, vlinc1010, vlinc1011, vlinc1012, and vlinc272, show little or no expression in HTD114 cells. Also shown in Figure 1 are the location of Fosmids used in the RNA-DNA FISH analyses (see below), the Long RNA-seq track (showing contiguous transcripts from three human cell lines, GM12878, HepG2, and K562) from ENCODE/Cold Spring Harbor, and the Repeat Masker Track showing the location of repetitive elements (also see Table S1).

Next, to determine if the vlinc273 transcripts show monoallelic expression in HTD114 cells, we used reverse transcribed RNA as input for PCR, followed by sequencing at heterozygous SNPs. Figure 2a shows sequencing traces from two different SNPs that are heterozygous in genomic DNA, but a single allele was detected in RNA isolated from HTD114 cells, indicating that these transcripts are monoallelic. In addition, we previously generated two chromosome 6 mono-chromosomal hybrids to aid in mapping heterozygous SNPs onto the HTD114 chromosome 6 homologs [6]. These two hybrid cell lines are mouse L cell clones, each containing one of the two chromosome 6s from

HTD114, which we arbitrarily name as CHR6A and CHR6B. Using these mono-chromosomal hybrids, we previously found that *ASAR6* is expressed from CHR6A [6]. Sequence traces generated from genomic DNA isolated from these mono-chromosomal hybrids indicated that the vlinc273 transcripts are derived from CHR6B (Fig. 2A), and therefore are expressed from the opposite chromosome, or in *trans*, to *ASAR6*.

We next assayed expression of the vlincRNA cluster using RNA-DNA FISH in HTD114 cells. For this analysis we used five different Fosmid probes to detect RNA (see Fig. 1), plus a chromosome 6 centromeric probe to detect DNA. As expected from the RNA-seq analysis, we did not detect expression from the genomic regions annotated as vlinc1012 and vlinc272 in HTD114 cells (not shown). In contrast, we detected expression of RNA, annotated as vinc273, that remains associated with one of the chromosome 6 homologs. Figure 2b-f show examples of this analysis using probes from within the vlinc273 locus to detect RNA. Note the relatively large clouds of RNA that are adjacent to, or overlapping with, one of the chromosome 6 centromeric DNA signals. In addition, we used RNA-DNA FISH to detect both ASAR6 and vlinc273 RNA in combination with a chromosome 6 whole chromosome paint as probe to detect chromosome 6 DNA. Figure 2g-j shows examples of this analysis and indicates that ASAR6 and vlinc273 RNAs are detected on opposite chromosome 6 homologs. We also note that the size of the RNA FISH signals detected by the ASAR6 and vlinc273 probes were variable, ranging from large clouds occupying the entire chromosome 6 territory, to relatively small spots of hybridization.

Random monoallelic expression of vlinc273:

The observations described above indicate that vlinc273 and ASAR6 RNAs are detected on opposite chromosome 6 homologs in the clonal cell line HTD114. This monoallelic expression could be due to either genomic imprinting, to DNA sequence polymorphisms within promoter or enhancer elements, or to PRME (see above). Therefore, to distinguish between these possibilities, we first determined if vlinc273 is monoallelically expressed in EBV transformed lymphoblasts, which have been used extensively in the analysis of autosomal monoallelic expression in humans [6, 8, 17, 20]. For this analysis, we used RNA-DNA FISH to assay expression of vlinc273 in GM12878 cells. For this analysis we used Fosmid probes to detect RNA (see Fig. 1), plus a chromosome 6 centromeric probe to detect DNA. We detected single sites of vlinc273 RNA hybridization in >95% of GM12878 cells (see Fig. 3A-E for examples).

Next, to determine if expression of vlinc273 is subject to PRME in human primary cells, we carried out RNA FISH on primary blood lymphocytes (PBLs) isolated from two unrelated individuals. For this analysis, we included an RNA FISH probe to the first intron of *KCNQ5*, which is a known PRME gene located on human chromosome 6 [6, 17, 20]. For this analysis, we used a two-color RNA FISH assay to detect expression of vlinc273 in combination with a probe from the first intron of *KCNQ5* on PBLs isolated from two unrelated individuals. Quantification of the number of RNA FISH signals in >100 cells indicated that vlinc273 and *KCNQ5* were expressed from the same chromosome 6 homolog in ~98% of cells from both individuals (see Fig. 3F-I for examples). Therefore, because *KCNQ5* expression is subject to PRME [6, 17, 20], and the PBLs are not clonally derived, we conclude that the monoallelic expression of vlinc273 must also be random and therefore not imprinted. We note that the size of the RNA hybridization signals

220 detected by the *KCNQ5* and vlinc273 probes were variable, ranging from large clouds to
221 relatively small sites of hybridization. We also detected two sites of hybridization for both
222 probes in ~2% of cells (Fig. 3J). Finally, to directly compare the appearance of the RNA
223 FISH signals detected for vlinc273 to XIST RNA expressed from the inactive X
224 chromosome we assayed vlinc273 and XIST RNAs simultaneously in female PBLs.
225 Figure 3k and 3i show the clouds of RNA detected by the vlinc273 probe in relation to the
226 relatively larger clouds of RNA hybridization detected by the *XIST* probe.
227
228 **Asynchronous replication of vlinc273 is coordinated on chromosome 6.**
229 All monoallelically expressed genes share the property of asynchronous replication
230 [27]. We previously used Replication Timing-Specific Hybridization (ReTiSH) [19] to assay
231 coordinated asynchronous replication of chromosome 6 loci, including *ASAR6*. In the
232 ReTiSH assay, cells are labeled with BrdU for different times and then harvested during
233 mitosis (see Fig. 4A). Regions of chromosomes that incorporate BrdU are visualized by a
234 modification of chromosome orientation-fluorescence in situ hybridization (CO-FISH),
235 where the replicated regions (BrdU-labeled) are converted to single stranded DNA and
236 then hybridized directly with specific probes [19]. Since mitotic chromosomes are analyzed
237 for hybridization signals located on the same chromosome in metaphase spreads, the
238 physical distance between the loci is not a limitation of the ReTiSH assay [19]. We
239 previously used this approach to show that the asynchronous replication of *ASAR6* was
240 coordinated in *cis* or in *trans* with other random monoallelic loci on human chromosome 6
241 [6, 20]. For this analysis, we used PBLs and a three-color hybridization scheme to
242 simultaneously detect the vlinc273 locus, *ASAR6*, and the chromosome 6 centromere. The

chromosome 6 centromeric probe was included to unambiguously identify both chromosome 6s. Because centromeric heterochromatin is late replicating, centromeric probes hybridize to both copies of chromosome 6 at the 14 and 5 hour time points [19]. We found that the vlinc273 alleles were subject to asynchronous replication that is coordinated in *cis* with *ASAR6* (Fig. 4B-D; and Table 1). Therefore, because the asynchronous replication of *ASAR6* is coordinated with other random monoallelic loci on chromosome 6 [6, 20], we conclude that the vlinc273 locus is part of a chromosome-wide system that coordinates the asynchronous replication of random monoallelic loci on chromosome 6.

One shared characteristic of the *ASAR6* and *ASAR15* genes is that the silent alleles replicate before the expressed alleles on their respective chromosomes [6, 7, 20]. Therefore, one unanticipated result from our ReTiSH assay is that asynchronous replication of vlinc273 and *ASAR6* is coordinated in *cis*. Thus, the earlier replicating vlinc273 allele is on the same homolog as the earlier replicating *ASAR6* allele. Therefore, to determine if the asynchronous replication of vlinc273 and *ASAR6* is also coordinated in *cis* in HTD114 cells, where they are expressed from opposite homologs (see Fig. 2), we analyzed the asynchronous replication of vlinc273 and *ASAR6* using the same three color ReTiSH assay describe above [19]. In addition, HTD114 cells contain a centromeric polymorphism on chromosome 6, and the chromosome with the larger centromere is linked to the later replicating and expressed allele of *ASAR6* [20]. We found that the asynchronous replication of vlinc273 and *ASAR6* is coordinated in *cis* in HTD114 cells (Fig. 4E-G; and Table 1). These observations are consistent with our previous finding that *ASAR6* is expressed from the later replicating allele ([20]; CHR6A), and indicate that

vlinc273 is expressed from the earlier replicating allele in HTD114 cells (CHR6B; see Fig. 2A). Regardless, we found that the vlinc273 locus is subject to random monoallelic expression and asynchronous replication that is coordinated with other random monoallelic loci on chromosome 6 and therefore vlinc273 is subject to PRME.

Deletion of the expressed allele of vlinc273 results in delayed replication in *cis*.

To determine if the genomic region containing the vlincRNA cluster located on chromosome 6 at 140.2-141.3 mb (see Fig. 1) regulates replication timing, we used CRISPR/Cas9 to delete the entire locus in HTD114 cells. For this analysis we designed single guide RNAs (sgRNAs) to unique sequences as shown in Fig 1. We expressed sgRNA-1 and sgRNA-3 in combination with Cas9 and screened clones for deletions using PCR primers that flank the sgRNA binding sites (see Fig. 1 and Table S2). Because vlinc273 expression is monoallelic in HTD114 cells (see Fig. 2), we isolated clones that had heterozygous deletions affecting either CHR6A or CHR6B. We determined which allele was deleted based on retention of the different base pairs of heterozygous SNPs located within the deleted regions (see Table S2).

From our previous studies, we knew that prior to any genetic alterations the chromosome 6 homologs replicate synchronously in HTD114 cells [5, 6, 20, 26]. In addition, we also took advantage of the centromeric polymorphism in HTD114 cells to unambiguously distinguish between the two chromosome 6 homologs ([20, 26]; see Fig. 4E-G). The chromosome 6 with the larger centromere is linked to the expressed allele of *ASAR6* ([20]; CHR6A), and therefore the expressed allele of vlinc273 is linked to the chromosome 6 with the smaller centromere (CHR6B). For this replication timing assay, cultures were incubated with BrdU for 5.5 hours and mitotic cells harvested, processed for

BrdU incorporation and subjected to FISH using a chromosome 6 centromeric probe. As expected, prior to disruption of the vlinc cluster, CHR6A and CHR6B display synchronous replication (see Fig. 5F below). In contrast, cells containing a deletion of the vlinc cluster on CHR6B contain significantly more BrdU incorporation into CHR6B than in CHR6A (Fig. 5A-E). Quantification of the BrdU incorporation in multiple cells indicated that deletion of the CHR6B allele, which contains the expressed allele of vlinc273, results in a significant delay in replication timing (Fig. 5F). This is in contrast to cells containing a deletion of the vlinc cluster from the CHR6A allele, which is silent for all 6 vlincRNAs, where the BrdU incorporation is comparable between CHR6A and CHR6B (Fig. 5F). In addition, replication timing analysis of heterozygous deletions encompassing only the vlinc273 locus (using sgRNA-2 and sgRNA-3) indicated that deletion of the expressed allele (CHR6B), but not the silent allele (CHR6A), resulted in delayed replication of chromosome 6 (Fig. 5F). Finally, deletion of the vlinc271, vlinc1010, vlinc1011, vlinc1012 and vlinc272 loci (using sgRNA-1 and sgRNA-2) on CHR6B, did not result in delayed replication of chromosome 6 (Fig. 5F). For an additional comparison, we included the chromosome 6 replication timing data from HTD114 cells containing heterozygous deletions of *ASAR6* on the expressed allele (CHR6A) and on the silent allele (CHR6B) (Fig. 5F; also see Fig. S1). Taken together these results indicate that deletion of the expressed allele of vlinc273 results in delayed replication of chromosome 6 in *cis*, and because vlinc273 also displays PRME, the vlinc273 locus is an ASAR. Because vlinc273 is the second ASAR identified on human chromosome 6 and is located at ~141 mb, we designate this gene as *ASAR6-141*.

Discussion

Chromosome associated lncRNAs have become well established as regulators of chromosome scale replication timing, gene expression and structural stability [1, 28]. In this report, we identified a second chromosome 6 lncRNA gene, *ASAR6-141*, that when disrupted results in delayed replication timing of the entire chromosome in *cis*. *ASAR6* and *ASAR6-141* are subject to PRME, are expressed from opposite chromosome 6 homologs, and disruption of the expressed alleles, but not the silent alleles, leads to delayed replication timing of human chromosome 6 in *cis*. ASAR6 and ASAR6-141 RNAs share certain characteristics, including RNA Pol II products that are non-spliced, non-polyadenylated, contain a high L1 content and remain associated with the chromosome territories where they are transcribed. Taken together our results indicate that the replication timing of human chromosome 6 is regulated by the reciprocal monoallelic expression of two different ASAR lncRNA genes (see Fig. 6).

We previously found that deletion of the expressed allele of *ASAR6* results in transcriptional activation of the previously silent alleles of other monoallelic genes nearby, indicating that *ASAR6* negatively regulates expression of the previously silent alleles of other linked monoallelic genes [6]. One important tool for the analysis of chromosome scale gene expression has been the use of Cot-1 DNA as an RNA FISH probe [29, 30]. Cot-1 DNA (which contains highly repetitive sequences) is routinely used to block non-specific hybridization of genomic probes to repeats, and has been developed as a probe to detect global gene expression using FISH [30]. Cot-1 RNA hybridization provides a convenient assay to identify silent heterochromatic regions within nuclei by the absence of a hybridization signal [30]. Our model for ASAR function predicts that ASAR lncRNAs expressed from every chromosome are detected by Cot-1 RNA FISH due to the presence

of repetitive sequences, including abundant L1 antisense sequences, within the ASAR transcripts [26]. This interpretation is consistent with the observation that Cot-1 RNA is comprised predominantly of L1 sequences and is associated with euchromatin throughout interphase nuclei [31]. Furthermore, L1 RNA is localized to interphase chromosome territories, is excluded from heterochromatin, and is associated with the euchromatin fraction of chromosomes even following prolonged transcriptional inhibition [31]. We previously found that ectopic integration of an *ASAR6* transgene leads to loss of Cot1 RNA on the integrated chromosome, suggesting that the ASAR6 transgene silenced the endogenous ASARs on the integrated chromosome [26]. Therefore, because *ASAR6* and *ASAR6-141* are expressed from opposite homologs, our model includes reciprocal silencing of each other in *cis,* resulting in the reciprocal pattern of monoallelic expression of *ASAR6* and *ASAR6-141* on the two chromosome 6 homologs (Fig. 6).

One hallmark of genes that are subject to PRME is coordination in the asynchronous replication between alleles [7-9, 20]. This coordination can be either in *cis*, i.e. the early replicating alleles of two genes are always on the same homolog; or in *trans*, i.e. the early replicating alleles are always on opposite homologs [20]. In this report, we found that the asynchronous replication of *ASAR6-141* is coordinated in *cis* with *ASAR6*. This observation is consistent with our previous findings that human chromosome 6 contains loci that display random asynchronous replication that is coordinated both in *cis* and in *trans*, that some of these asynchronous loci are separated by >100 megabases of genomic DNA, and that the coordinated loci are on either side of the centromere [6, 20]. It will be interesting to determine if all human autosome pairs display a similar coordination in expression and asynchronous replication of PRME genes.

Asynchronous replication of random monoallelic genes is an epigenetic mark that appears before transcription and is thought to underlie the differential expression of the two alleles of identical sequence [18]. Therefore, because the asynchronous replication at PRME genes is coordinated along each chromosome, the expression pattern of PRME genes is also anticipated to be coordinated, i.e. in *cis*- always expressed from the same homolog; or in *trans*- always expressed from opposite homologs. We previously found that *ASAR6* and *ASAR15* are expressed from the later replicating alleles [7, 20]. In contrast, the *FUT9* protein coding gene, which is closely linked to *ASAR6* (see Fig. S2), is expressed from the early replicating allele [7, 20]. Therefore, PRME genes can be expressed from either the early or the late replicating alleles. One unanticipated result from our allelic expression and asynchronous replication assays described here is that *ASAR6-141* is expressed from the early replicating allele, which is the first example of an ASAR that is expressed from the early replicating allele. Nevertheless, we found that disruption of the expressed allele, but not the silent allele of *ASAR6-141* results in delayed replication of chromosome 6, indicating that expression and not asynchronous replication is a critical component of ASAR function. This conclusion is consistent with our previous observation that ASAR6 RNA mediates the chromosome-wide effects of *ASAR6* forced expression [26]. Therefore, the role of asynchronous replication at ASAR loci may serve as a mechanism to help establish which allele will be transcribed. Thus, the epigenetic mark that establishes early and late replication between the two alleles of PRME genes may function to establish asymmetry between alleles, and then depending on the promoter/enhancer elements at different PRME genes either the early or late replicating allele will be transcribed.

One striking feature of both *ASAR6* and *ASAR15* is that they contain a high density L1 retrotransposons, constituting ~40% and ~55% of the expressed sequence, respectively [6, 7]. L1s were first implicated in monoallelic expression when Dr. Mary Lyon proposed that L1s represent "booster elements" that function during the spreading of X chromosome inactivation [32, 33]. In humans, the X chromosome contains ~27% L1 derived sequence while autosomes contain ~13% [34]. In addition, L1s are present at a lower concentration in regions of the X chromosome that escape inactivation, supporting the hypothesis that L1s serve as signals to propagate inactivation along the X chromosome [34]. Further support for a role of L1s in monoallelic expression came from the observation that L1s are present at a relatively high local concentration near both imprinted and random monoallelic genes located on autosomes [35]. L1s have also been linked to DNA replication timing from the observation that differentiation-induced replication timing changes are restricted to AT rich isochores containing high L1 density [36]. Another potential link between L1s and DNA replication is the observation that ~25% of origins in the human genome were mapped to L1 sequences [37]. While this observation is suggestive of a relationship between origins and L1s, it is not clear what distinguishes L1s with origin activity from L1s without [37].

During our genetic characterization of *ASAR6* we mapped an ~29 kb critical region that when deleted results in DRT/DMC [6, 20]. This ~29 kb region contains one full length and 5 truncated L1s. Similarly, we mapped an ~124 kb critical region within *ASAR15* that contains 3 full length and 15 truncated L1s [7]. We recently used ectopic integration of transgenes and CRISPR/Cas9-mediated chromosome engineering and found that L1 sequences, oriented in the antisense direction, mediate the chromosome-wide effects of

405 *ASAR6* and *ASAR15* [38]. In addition, we found that oligonucleotides targeting the
406 antisense strand of the one full length L1 within ASAR6 RNA restored normal replication
407 timing to mouse chromosomes expressing an *ASAR6* transgene. These results provided
408 the first direct evidence that L1 antisense RNA plays a functional role in replication timing
409 of mammalian chromosomes [38].

410 We previously proposed a model in which the antisense L1 sequences function to
411 suppress splicing, and to promote stable association of the RNA with the chromosome
412 territories where they are transcribed [26]. Consistent with this interpretation is the finding
413 that a de novo L1 insertion, in the antisense orientation, into an exon of the mouse *Nr2e3*
414 gene results in inefficient splicing, accumulation of the transcript to high levels, and
415 retention of the transcript at the mutant *Nr2e3* locus [39]. In addition, a more recent study
416 found that the antisense strand of L1 RNA functions as a multivalent "hub" for binding to
417 numerous nuclear matrix and RNA processing proteins, and that the L1 antisense RNA
418 binding proteins repress splicing and 3' end processing within and around the L1s [40].

419 The vlincRNAs were identified as RNA transcripts of >50 kb of contiguous RNA-seq
420 reads that have no overlap with annotated protein coding genes [23]. The vlincRNAs were
421 identified from the FANTOM5 Cap Analysis of Gene Expression (CAGE) dataset,
422 indicating that the vlincRNAs contain 5' caps and consequently represent RNA Pol II
423 transcripts [23]. We previously found that *ASAR6* and *ASAR15* are also transcribed by
424 RNA Pol II [6, 7]. ASAR6-141 RNA shares certain characteristics with ASAR6 and
425 ASAR15 RNAs that distinguish them from other canonical RNA Pol II lncRNAs. Thus,
426 even though ASAR6-141, ASAR6 and ASAR15 RNAs are RNA Pol II products they show
427 little or no evidence of splicing or polyadenylation and remain associated with the

chromosome territories where they were transcribed ([6, 7, 23]; and see Fig. 2 and S2). Our work supports a model where all mammalian chromosomes express "ASAR" genes that encode chromosome associated lncRNAs that control the replication timing program in *cis*. In this model, the ASAR lncRNAs function to promote proper chromosome replication timing by controlling the timing of origin firing. In addition, because both *ASAR6,* and *ASAR6-141* are monoallelically expressed, our model includes expression of different ASAR genes from opposite homologs ([26]; see Fig. 6).

We previously found that 5% of chromosome translocations, induced by two different mechanisms (IR and Cre/loxP) display DRT/DMC [4, 5]. Because ~5% of translocations display DRT/DMC and only one of the two translocation products has DRT/DMC [5], these results indicate that ~2.5% of translocation products display DRT/DMC [4-7]. Taken with the observation that the translocations that display DRT/DMC have disrupted ASAR genes [6, 7], suggests that ~2.5% of the genome is occupied by ASARs. The vlincRNAs were identified as nuclear, non-spliced, non-polyadenylated transcripts of >50 kb of contiguously expressed sequence that are not associated with protein coding genes [23]. There are currently >2,700 annotated human vlincRNAs, and they are expressed in a highly cell type-specific manner [23-25]. Because many of the vlincRNAs are encoded by regions of the genome that do not overlap with protein coding genes, many of the vlincRNAs contain a high density of repetitive elements, including L1s (see Fig. 1 and Table S1 for examples). In this report, we found that the genomic region annotated as vlinc273 has all of the physical and functional characteristics that are shared with *ASAR6* and *ASAR15*, and therefore vlinc273 is an ASAR (designated here as *ASAR6-141*). In addition, while ASAR6 RNA was not annotated as a vlincRNA in any previous publication,

451 our RNA-seq data from HTD114 cells indicates that ASAR6 RNA has all of the
452 characteristics of a vlincRNA (see Fig. S2). Furthermore, we note that there are two
453 annotated vlincRNAs (vlinc253 and vlinc254) that map within the ~1.2 mb domain of *cis*-
454 coordinated asynchronous replication that we previously associated with the *ASAR6*
455 locus ([20]; Fig. S2). Therefore, vlinc253 and vlinc254 display asynchronous replication
456 that is coordinated with *ASAR6*, *ASAR6-141* and all other PRME genes on human
457 chromosome 6 (see [20]). Taken together, these observations raise the intriguing
458 possibility that these other vlincRNAs are also ASARs. Finally, the clustering of vlincRNA
459 genes with ASAR characteristics, and their apparent tissue-restricted expression patterns
460 (see Fig. 1 and S2), supports a model in which each autosome contains clustered ASAR
461 genes, and that these ASAR clusters, expressing different ASAR transcripts in different
462 tissues, function as “Inactivation/Stability Centers” that control replication timing,
463 monoallelic gene expression, and structural stability of each chromosome.
464

Methods:

Cell culture

HTD114 cells are a human *APRT* deficient cell line derived from HT1080 cells [41], and were grown in DMEM (Gibco) supplemented with 10% fetal bovine serum (Hyclone). GM12878 cells were obtained from ATCC and were grown in RPMI 1640 (Life Technologies) supplemented with 15% fetal bovine serum (Hyclone). Primary blood lymphocytes were isolated after venipuncture into a Vacutainer CPT (Becton Dickinson, Franklin Lakes, NJ) per the manufacturer's recommendations and grown in 5 mL RPMI 1640 (Life Technologies) supplemented with 10% fetal bovine serum (Hyclone) and 1% phytohemagglutinin (Life Technologies). All cells were grown in a humidified incubator at 37°C in a 5% carbon dioxide atmosphere.

RNA-seq:

Nuclei were isolated from HTD114 cells following lysis in 0.5% NP40, 140 mM NaCl, 10 mM Tris-HCl (pH 7.4), and 1.5 mM $MgCl_2$. Nuclear RNA was isolated using Trizol reagent using the manufacturer's instructions, followed by DNase treatment to remove possible genomic DNA contamination. RNA-seq was carried out at Novogene. Briefly, ribosomal RNAs were removed using the Ribo-Zero kit (Illumina), RNA was fragmented into 250-300bp fragments, and cDNA libraries were prepared using the Directional RNA Library Prep Kit (NEB). Paired end sequencing was done on a NoaSeq 6000. Triplicate samples were merged and aligned to the human genome (hg19) using the STAR aligner [42] with default settings. Duplicate reads and reads with map quality below 30 were removed with SAMtools [43].

488

489 **DNA FISH**

490 Mitotic chromosome spreads were prepared as described previously [2]. After RNase
491 (100µg/ml) treatment for 1h at 37˚C, slides were washed in 2XSSC and dehydrated in an
492 ethanol series and allowed to air dry. Chromosomal DNA on the slides was denatured at
493 75˚C for 3 minutes in 70% formamide/2XSSC, followed by dehydration in an ice cold
494 ethanol series and allowed to air dry. BAC and Fosmid DNAs were labeled using nick
495 translation (Vysis, Abbott Laboratories) with Spectrum Orange-dUTP, Spectrum Aqua-
496 dUTP or Spectrum Green-dUTP (Vysis). Final probe concentrations varied from 40-60
497 ng/µl. Centromeric probe cocktails (Vysis) and/or whole chromosome paint probes
498 (Metasystems) plus BAC or Fosmid DNAs were denatured at 75˚C for 10 minutes and
499 prehybridized at 37˚C for 10 minutes. Probes were applied to denatured slides and
500 incubated overnight at 37˚C. Post-hybridization washes consisted of one 3-minute wash
501 in 50% formamide/2XSSC at 40˚C followed by one 2-minute rinse in PN (0.1M Na_2HPO_4,
502 pH 8.0/2.5% Nonidet NP-40) buffer at RT. Coverslips were mounted with Prolong Gold
503 antifade plus DAPI (Invitrogen) and viewed under UV fluorescence (Olympus).

504

505 **ReTiSH**

506 We used the ReTiSH assay essentially as described [19]. Briefly, unsynchronized,
507 exponentially growing cells were treated with 30µM BrdU (Sigma) for 6 or 5 and 14 hours.
508 Colcemid (Sigma) was added to a final concentration of 0.1 µg/mL for 1 h at 37°C. Cells
509 were trypsinized, pelleted by centrifugation at 1,000 rpm, and resuspended in prewarmed
510 hypotonic KCl solution (0.075 M) for 40 min at 37°C. Cells were pelleted by centrifugation

2:

and fixed with methanol-glacial acetic acid (3:1). Fixed cells were drop gently onto wet, cold slides and allowed to air-dry. Slides were treated with 100µg/ml RNAse A at 37°C for 10 min. Slides were rinsed briefly in H_20 followed by fixation in 4% formaldehyde at room temperature for 10 minutes. Slides were incubated with pepsin (1 mg/mL in 2N HCl) for 10 min at 37°C, and then rinsed again with H_20 and stained with 0.5 µg/µL Hoechst 33258 (Sigma) for 15 minutes. Slides were flooded with 200µl 2xSSC, coversliped and exposed to 365-nm UV light for 30 min using a UV Stratalinker 2400 transilluminator (Stratagene). Slides were rinsed with H_20 and drained. Slides were incubated with 100µl of 3U/µl of ExoIII (Fermentas) in ExoIII buffer for 15 min at 37°C. The slides were then processed directly for DNA FISH as described above, except with the absence of a denaturation step. *ASAR6* DNA was detected with BAC RP11-767E7, and *ASAR6-141* DNA was detected with BAC RP11-715D3.

RNA-DNA FISH

Cells were plated on glass microscope slides at ~50% confluence and incubated for 4 hours in complete media in a 37ºC humidified CO_2 incubator. Slides were rinsed 1X with sterile RNase free PBS. Cell Extraction was carried out using ice cold solutions as follows: Slides were incubated for 30 seconds in CSK buffer (100mM NaCl/300mM sucrose/3mM $MgCl_2$/10mM PIPES, pH 6.8), 10 minutes in CSK buffer/0.1% Triton X-100, followed by 30 seconds in CSK buffer. Cells were then fixed in 4% paraformaldehyde in PBS for 10 minutes and stored in 70% EtOH at -20°C until use. Just prior to RNA FISH, slides were dehydrated through an EtOH series and allowed to air dry. Denatured probes were prehybridized at 37°C for 10 min, applied to non-denatured slides and hybridized at 37ºC

for 14-16 hours. Post-hybridization washes consisted of one 3-minute wash in 50% formamide/2XSSC at 40˚C followed by one 2-minute rinse in 2XSSC/0.1% TX-100 for 1 minute at RT. Slides were then fixed in 4% paraformaldehyde in PBS for 5 minutes at RT, and briefly rinsed in 2XSSC/0.1% TX-100 at RT. Coverslips were mounted with Prolong Gold antifade plus DAPI (Invitrogen) and slides were viewed under UV fluorescence (Olympus). Z-stack images were generated using a Cytovision workstation. After capturing RNA FISH signals, the coverslips were removed, the slides were dehydrated in an ethanol series, and then processed for DNA FISH, beginning with the RNase treatment step, as described above.

Replication timing assay

The BrdU replication timing assay was performed as described previously on exponentially dividing cultures and asynchronously growing cells [44]. Mitotic chromosome spreads were prepared and DNA FISH was performed as described above. The incorporated BrdU was then detected using a FITC-labeled anti-BrdU antibody (Roche). Coverslips were mounted with Prolong Gold antifade plus DAPI (Invitrogen), and viewed under UV fluorescence. All images were captured with an Olympus BX Fluorescent Microscope using a 100X objective, automatic filter-wheel and Cytovision workstation. Individual chromosomes were identified with either chromosome-specific paints, centromeric probes, BACs or by inverted DAPI staining. Utilizing the Cytovision workstation, each chromosome was isolated from the metaphase spread and a line drawn along the middle of the entire length of the chromosome. The Cytovision software was used to calculate the pixel area and intensity along each chromosome for each

fluorochrome occupied by the DAPI and BrdU (FITC) signals. The total amount of fluorescent signal in each chromosome was calculated by multiplying the average pixel intensity by the area occupied by those pixels. The BrdU incorporation into human chromosome 6 homologs containing CRISPR/Cas9 modifications was calculated by dividing the total incorporation into the chromosome with the smaller chromosome 6 centromere (6B) divided by the BrdU incorporation into the chromosome 6 with the larger centromere (6A) within the same cell. Boxplots were generated from data collected from 8-12 cells per clone or treatment group. Differences in measurements were tested across categorical groupings by using the Kruskal-Wallis test [45] and listed as P-values for the corresponding plots.

CRISPR/Cas9 engineering

Using Lipofectamine 2000, according to the manufacturer's recommendations, we co-transfected HTD114 cells with plasmids encoding GFP, sgRNAs and Cas9 endonuclease (Origene). Each plasmid encoded sgRNAs were designed to bind at the indicated locations (Fig. 1; also see Table S1). 48h after transfection, cells were plated at clonal density and allowed to expand for 2-3 weeks. The presence of deletions in were confirmed by PCR using the primers described in Supplemental Table S1. The single cell colonies that grew were analyzed for heterozygous deletions by PCR. We used retention of a heterozygous SNPs (see Table S1) to identify the disrupted allele (CHR6A vs CHR6B), and homozygosity at this SNP confirmed that cell clones were homogenous.

References

1. Thayer MJ. Mammalian chromosomes contain cis-acting elements that control replication timing, mitotic condensation, and stability of entire chromosomes. Bioessays. 2012;34(9):760-70. PubMed PMID: 22706734.

2. Smith L, Plug A, Thayer M. Delayed Replication Timing Leads to Delayed Mitotic Chromosome Condensation and Chromosomal Instability of Chromosome Translocations. Proc Natl Acad Sci U S A. 2001;98:13300-5.

3. Chang BH, Smith L, Huang J, Thayer M. Chromosomes with delayed replication timing lead to checkpoint activation, delayed recruitment of Aurora B and chromosome instability. Oncogene. 2007;26(13):1852-61. PubMed PMID: 17001311.

4. Breger KS, Smith L, Turker MS, Thayer MJ. Ionizing radiation induces frequent translocations with delayed replication and condensation. Cancer Research. 2004;64:8231-8.

5. Breger KS, Smith L, Thayer MJ. Engineering translocations with delayed replication: evidence for cis control of chromosome replication timing. Hum Mol Genet. 2005;14(19):2813-27. PubMed PMID: 16115817.

6. Stoffregen EP, Donley N, Stauffer D, Smith L, Thayer MJ. An autosomal locus that controls chromosome-wide replication timing and mono-allelic expression. Hum Mol Genet. 2011;20:2366-78. PubMed PMID: 21459774.

7. Donley N, Smith L, Thayer MJ. ASAR15, A cis-Acting Locus that Controls Chromosome-Wide Replication Timing and Stability of Human Chromosome 15.

PLoS Genet. 2015;11(1):e1004923. Epub 2015/01/09. doi: 10.1371/journal.pgen.1004923. PubMed PMID: 25569254.

8. Ensminger AW, Chess A. Coordinated replication timing of monoallelically expressed genes along human autosomes. Hum Mol Genet. 2004;13(6):651-8. PubMed PMID: 14734625.

9. Singh N, Ebrahimi FA, Gimelbrant AA, Ensminger AW, Tackett MR, Qi P, et al. Coordination of the random asynchronous replication of autosomal loci. Nat Genet. 2003;33(3):339-41. PubMed PMID: 12577058.

10. Gendrel AV, Marion-Poll L, Katoh K, Heard E. Random monoallelic expression of genes on autosomes: Parallels with X-chromosome inactivation. Semin Cell Dev Biol. 2016. Epub 2016/04/23. doi: 10.1016/j.semcdb.2016.04.007. PubMed PMID: 27101886.

11. Gendrel AV, Attia M, Chen CJ, Diabangouaya P, Servant N, Barillot E, et al. Developmental dynamics and disease potential of random monoallelic gene expression. Dev Cell. 2014;28(4):366-80. Epub 2014/03/01. doi: 10.1016/j.devcel.2014.01.016. PubMed PMID: 24576422.

12. Chess A. Mechanisms and consequences of widespread random monoallelic expression. Nat Rev Genet. 2012;13(6):421-8. PubMed PMID: 22585065.

13. Alexander MK, Mlynarczyk-Evans S, Royce-Tolland M, Plocik A, Kalantry S, Magnuson T, et al. Differences between homologous alleles of olfactory receptor genes require the Polycomb Group protein Eed. J Cell Biol. 2007;179(2):269-76. PubMed PMID: 17954609.

626 14. Li SM, Valo Z, Wang J, Gao H, Bowers CW, Singer-Sam J.
627 Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic
628 expression within novel gene families. PLoS One. 2012;7(2):e31751. PubMed
629 PMID: 22384067.
630 15. Lin M, Hrabovsky A, Pedrosa E, Wang T, Zheng D, Lachman HM. Allele-
631 biased expression in differentiating human neurons: implications for
632 neuropsychiatric disorders. PLoS One. 2012;7(8):e44017. PubMed PMID:
633 22952857.
634 16. Bartolomei MS. Genomic imprinting: employing and avoiding epigenetic
635 processes. Genes Dev. 2009;23(18):2124-33. PubMed PMID: 19759261.
636 17. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread
637 monoallelic expression on human autosomes. Science. 2007;318(5853):1136-40.
638 PubMed PMID: 18006746.
639 18. Mostoslavsky R, Singh N, Tenzen T, Goldmit M, Gabay C, Elizur S, et al.
640 Asynchronous replication and allelic exclusion in the immune system. Nature.
641 2001;414(6860):221-5. PubMed PMID: 11700561.
642 19. Schlesinger S, Selig S, Bergman Y, Cedar H. Allelic inactivation of rDNA
643 loci. Genes Dev. 2009;23(20):2437-47. PubMed PMID: 19833769.
644 20. Donley N, Stoffregen EP, Smith L, Montagna C, Thayer MJ.
645 Asynchronous Replication, Mono-Allelic Expression, and Long Range Cis-Effects
646 of ASAR6. PLoS Genet. 2013;9(4):e1003423. PubMed PMID: 23593023.
647 21. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH,
648 et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is

649 'dark matter' un-annotated RNA. BMC biology. 2010;8:149. Epub 2010/12/24.
650 doi: 10.1186/1741-7007-8-149. PubMed PMID: 21176148; PubMed Central
651 PMCID: PMCPMC3022773.
652 22. St Laurent G, Savva YA, Kapranov P. Dark matter RNA: an intelligent
653 scaffold for the dynamic regulation of the nuclear information landscape. Front
654 Genet. 2012;3:57. Epub 2012/04/28. doi: 10.3389/fgene.2012.00057. PubMed
655 PMID: 22539933; PubMed Central PMCID: PMCPMC3336093.
656 23. St Laurent G, Vyatkin Y, Antonets D, Ri M, Qi Y, Saik O, et al. Functional
657 annotation of the vlinc class of non-coding RNAs using systems biology
658 approach. Nucleic Acids Res. 2016;44(7):3233-52. Epub 2016/03/24. doi:
659 10.1093/nar/gkw162. PubMed PMID: 27001520; PubMed Central PMCID:
660 PMCPMC4838384.
661 24. St Laurent G, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, et al.
662 VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and
663 cancer. Genome Biol. 2013;14(7):R73. Epub 2013/07/24. doi: 10.1186/gb-2013-
664 14-7-r73. PubMed PMID: 23876380; PubMed Central PMCID:
665 PMCPMC4053963.
666 25. Caron M, St-Onge P, Drouin S, Richer C, Sontag T, Busche S, et al. Very
667 long intergenic non-coding RNA transcripts and expression profiles are
668 associated to specific childhood acute lymphoblastic leukemia subtypes. PLoS
669 One. 2018;13(11):e0207250. Epub 2018/11/16. doi:
670 10.1371/journal.pone.0207250. PubMed PMID: 30440012; PubMed Central
671 PMCID: PMCPMC6237371.

672 26. Platt EJ, Smith L, Thayer MJ. L1 retrotransposon antisense RNA within
673 ASAR lncRNAs controls chromosome-wide replication timing. J Cell Biol. 2017.
674 Epub 2017/12/31. doi: 10.1083/jcb.201707082. PubMed PMID: 29288153.
675 27. Goldmit M, Bergman Y. Monoallelic gene expression: a repertoire of
676 recurrent themes. Immunol Rev. 2004;200:197-214. PubMed PMID: 15242406.
677 28. Galupa R, Heard E. X-Chromosome Inactivation: A Crossroads Between
678 Chromosome Architecture and Gene Regulation. Annu Rev Genet. 2018;52:535-
679 66. Epub 2018/09/27. doi: 10.1146/annurev-genet-120116-024611. PubMed
680 PMID: 30256677.
681 29. Creamer KM, Lawrence JB. XIST RNA: a window into the broader role of
682 RNA in nuclear chromosome architecture. Philosophical transactions of the
683 Royal Society of London Series B, Biological sciences. 2017;372(1733). Epub
684 2017/09/28. doi: 10.1098/rstb.2016.0360. PubMed PMID: 28947659; PubMed
685 Central PMCID: PMCPMC5627162.
686 30. Hall LL, Byron M, Sakai K, Carrel L, Willard HF, Lawrence JB. An ectopic
687 human XIST gene can induce chromosome inactivation in postdifferentiation
688 human HT-1080 cells. Proc Natl Acad Sci U S A. 2002;99(13):8677-82. PubMed
689 PMID: 12072569.
690 31. Hall LL, Carone DM, Gomez AV, Kolpa HJ, Byron M, Mehta N, et al.
691 Stable C0T-1 repeat RNA is abundant and is associated with euchromatic
692 interphase chromosomes. Cell. 2014;156(5):907-19. Epub 2014/03/04. doi:
693 10.1016/j.cell.2014.01.042. PubMed PMID: 24581492; PubMed Central PMCID:
694 PMCPmc4023122.

695 32. Lyon MF. X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell
696 Genet. 1998;80(1-4):133-7. PubMed PMID: 9678347.
697 33. Lyon MF. The Lyon and the LINE hypothesis. Semin Cell Dev Biol.
698 2003;14(6):313-8. PubMed PMID: 15015738.
699 34. Bailey JA, Carrel L, Chakravarti A, Eichler EE. Molecular evidence for a
700 relationship between LINE-1 elements and X chromosome inactivation: the Lyon
701 repeat hypothesis. Proc Natl Acad Sci U S A. 2000;97(12):6634-9. Epub
702 2000/06/07. PubMed PMID: 10841562; PubMed Central PMCID:
703 PMCPMC18684.
704 35. Allen E, Horvath S, Tong F, Kraft P, Spiteri E, Riggs AD, et al. High
705 concentrations of long interspersed nuclear element sequence distinguish
706 monoallelically expressed genes. Proc Natl Acad Sci U S A. 2003;100(17):9940-
707 5. PubMed PMID: 12909712.
708 36. Hiratani I, Leskovar A, Gilbert DM. Differentiation-induced replication-
709 timing changes are restricted to AT-rich/long interspersed nuclear element
710 (LINE)-rich isochores. Proc Natl Acad Sci U S A. 2004;101(48):16861-6. Epub
711 2004/11/24. doi: 10.1073/pnas.0406687101. PubMed PMID: 15557005; PubMed
712 Central PMCID: PMCPMC534734.
713 37. Bartholdy B, Mukhopadhyay R, Lajugie J, Aladjem MI, Bouhassira EE.
714 Allele-specific analysis of DNA replication origins in mammalian cells. Nat
715 Commun. 2015;6:7051. Epub 2015/05/20. doi: 10.1038/ncomms8051. PubMed
716 PMID: 25987481; PubMed Central PMCID: PMCPMC4479011.

3

717 38. Platt EJ, Smith L, Thayer MJ. L1 retrotransposon antisense RNA within
718 ASAR lncRNA genes controls chromosome-wide replication timing. Journal of
719 Cell Biology. 2017;in press. Epub December 29, 2017.
720 39. Chen J, Rattner A, Nathans J. Effects of L1 retrotransposon insertion on
721 transcript processing, localization and accumulation: lessons from the retinal
722 degeneration 7 mouse and implications for the genomic ecology of L1 elements.
723 Hum Mol Genet. 2006;15(13):2146-56. Epub 2006/05/26. doi:
724 10.1093/hmg/ddl138. PubMed PMID: 16723373.
725 40. Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, et
726 al. Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA
727 Processing. Cell. 2018;174(5):1067-81.e17. Epub 2018/08/07. doi:
728 10.1016/j.cell.2018.07.001. PubMed PMID: 30078707; PubMed Central PMCID:
729 PMCPMC6108849.
730 41. Zhu Y, Bye S, Stambrook PJ, Tischfield JA. Single-base deletion induced
731 by benzo[a]pyrene diol epoxide at the adenine phosphoribosyltransferase locus
732 in human fibrosarcoma cell lines. Mutat Res. 1994;321(1-2):73-9.
733 42. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.
734 STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England).
735 2013;29(1):15-21. Epub 2012/10/30. doi: 10.1093/bioinformatics/bts635. PubMed
736 PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.
737 43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The
738 Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford,
739 England). 2009;25(16):2078-9. Epub 2009/06/10. doi:

3:

10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.

44. Smith L, Thayer M. Chromosome replicating timing combined with fluorescent in situ hybridization. J Vis Exp. 2012;(70):e4400. Epub 2012/12/29. doi: 10.3791/4400. PubMed PMID: 23271586; PubMed Central PMCID: PMCPMC3567166.

45. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29:1–27.

Figure Legends

Fig 1. UCSC Genome Browser view of the vlinc cluster on chromosome 6 between 140.3 and 141.3 mb. The genomic locations of vlinc271, vlinc1010, vlinc1011, vlinc1012, vlinc272 and vlinc273 are illustrated using the UCSC Genome Browser. RNA-seq data from nuclear RNA isolated from HTD114 is shown (HTD114 RNA-seq). Long RNA-seq data from the ENCODE Project (Cold Spring Harbor Lab) is shown using the Contigs view. Expression from the human cell lines GM12878 (red), K562 (dark blue), Hela S3 (light blue), and HepG2 (magenta) are shown. RNA from total cellular Poly A+ (cel pA+), total cellular Poly A- (cel pA-), nuclear Poly A+ (nuc pA+), nuclear Poly A- (nuc pA-), cytoplasmic Poly A+ (cyt pA+), and cytoplasmic Poly A- (cyt pA-) are shown. Also shown are the Repeating Elements using the RepeatMasker track. The location of five RNA FISH probes (Fosmids) that were used to detect expression of vlinc1012 (G248P87615G6), vlinc272 (G248P89033E3 and G248P83206B11) and vlinc273 (G248P85904G6 and G248P81345F10) are shown.

Fig 2. Mono-allelic expression and nuclear retention of vlinc273 in HTD114 cells. (A) DNA sequencing traces from PCR products designed to detect SNPs rs989613623 and rs2328092. PCRs were carried out on genomic DNAs isolated from HTD114, two mono-chromosomal hybrids containing the two different chromosome 6s from HTD114 {L(Hyg)-1 contains chromosome 6A (CHR6A) and expresses *ASAR6*, and L(Neo)-38 contains chromosome 6B (CHR6B) and is silent for *ASAR6* [6]}. The top and bottom panels also show the traces from HTD114 cDNA (RNA). The asterisks mark the location of the

773 heterozygous SNPs. (B-F) RNA-DNA FISH to detect vlinc273 expression in HTD114
774 cells. Fosmid G248P81345F10 was used as probe to detect vlinc273 RNA (green), and
775 a chromosome 6 centromeric probe was used to detect chromosome 6 DNA (red). The
776 nuclear DNA was stained with DAPI. Bars are 2.5 uM. (G-J) RNA-DNA FISH to detect
777 vlinc273 and ASAR6 expression in HTD114 cells. Fosmid G248P81345F10 was used as
778 probe to detect vlinc273 RNA (green), Fosmid G248P86031A6 was used as probe to
779 detect ASAR6 RNA (red), and a chromosome 6 paint was used to detect chromosome 6
780 DNA (magenta). The nuclear DNA was stained with DAPI. Bars are 2.5 uM.
781
782 **Fig 3.** Mono-allelic expression and nuclear retention of vlinc273 in EBV transformed
783 lymphoblasts and primary blood lymphocytes. (A-E) RNA-DNA FISH to detect vlinc273
784 expression in GM12878 EBV transformed lymphocytes. Fosmid G248P81345F10 was
785 used to detect vlinc273 RNA (green), and a chromosome 6 centromeric probe (CHR6
786 cen) was used to detect chromosome 6 DNA (red). (F-J) RNA FISH to detect coordinated
787 expression of vlinc273 and KCNQ5, a known random monoallelic gene, in primary blood
788 lymphocytes. Fosmid G248P81345F10 was used to detect vlinc273 RNA (green), and
789 Fosmid G248P80791F6 was used to detect expression of the first intron of KCNQ5. (K
790 and l) RNA FISH to detect expression of vlinc273 and XIST RNAs, in female primary
791 blood lymphocytes. The nuclear DNA was stained with DAPI. Bars are 2.5 uM.
792
793 **Fig 4.** Coordinated asynchronous replication timing on chromosome 6. (A) Schematic
794 representation of the ReTiSH assay. Cells were exposed to BrdU during the entire length
795 of S phase (14 hours) or only during late S phase (5 hours). The ReTiSH assay can

distinguish between alleles that replicate early (E) and late (L) in S phase. (B-D) Mitotic spreads that were processed for ReTiSH were hybridized with three different FISH probes. First, each hybridization included a centromeric probe to chromosome 6 (magenta). Arrows mark the centromeric signals in panels B (5 hours) and C (14 hours). Each assay also included BAC probes representing *ASAR6* (RP11-374I15; green) and vlinc273 (RP11-715D3; red). Panel D shows the two chromosome 6s, from both the 5 hour and 14 hour time points, aligned at their centromeres. The *ASAR6* BAC and the vlinc273 BAC show hybridization signals on the same chromosome 6 at the 5-hour time point, and as expected hybridized to both chromosome 6s at the 14 hour time point. The chromosomal DNA was stained with DAPI. (E-G) ReTiSH assay on HTD114 cells. Each ReTiSH assay included a centromeric probe to chromosome 6 (magenta). Arrows mark the centromeric signals in panels e (5 hours) and f (14 hours). Each assay also included BAC probes for *ASAR6* (RP11-374I15; green) and *vlinc273* (RP11-715D3; red). The chromosomal DNA was stained with DAPI. The *ASAR6* BAC and the vlinc273 BAC show hybridization signals to the same chromosome 6s (CHR6A) at the 5-hour time point, and as expected hybridized to both chromosome 6s at the 14-hour time point.

Fig 5. Delayed replication of chromosome 6 following disruption of vlinc273. (A and B) A representative mitotic spread from BrdU (green) treated cells containing a deletion of the expressed allele of the vlinc273 locus. Mitotic cells were subjected to DNA FISH using a chromosome 6 centromeric probe (red). The larger centromere resides on the chromosome 6 with the expressed *ASAR6* allele and the silent vlinc273 allele (6A). (C) The two chromosome 6s were extracted from a and b and aligned to show the BrdU

3'

819 incorporation and centromeric signals. (D) Pixel intensity profiles of BrdU incorporation
820 and DAPI staining along the (6A) and (6B) chromosomes from panel C. (E) BrdU
821 quantification along 6A and 6B from panel D. (F) The ratio of DNA synthesis into the two
822 chromosome 6s was calculated by dividing the BrdU incorporation in 6B by the
823 incorporation in 6A in multiple cells. The box plots show the ratio of incorporation before
824 (Intact, dark blue), and heterozygous deletions of the entire locus (Δ6A purple; and Δ6B
825 light blue), which included vlinc271, vlinc1010, vlinc1011, vlinc1012, vlinc272, and
826 vlinc273, see map in Fig 1. Heterozygous deletions affecting vlinc273 only from the silent
827 (Δ6A orange) or expressed (Δ6B green) alleles are shown. A heterozygous deletion
828 affecting vlinc271, vlinc1010, vlinc1011, vlinc1012, and vlinc272 on CHR6B (Δ6B) is
829 shown in pink. Also shown are the heterozygous deletions affecting *ASAR6* from on the
830 expressed (Δ6A magenta) or silent (Δ6B yellow) alleles. P values of $<1 \times 10^{-4}$ are indicated
831 by ***, and P values of $>1 \times 10^{-1}$ are indicated by *, and were calculated using the Kruskal-
832 Wallis test. Error bars are SD.
833
834 **Fig 6. "ASAR" model of replication timing on chromosome 6.** The two homologs of
835 human chromosome 6 are shown (gray) with origins of replication depicted as blue bars.
836 Expression of *ASAR6* and *ASAR6-141* genes is monoallelic, resulting in a reciprocal
837 expression pattern with an expressed or active ASAR (green or red clock) and a silent or
838 inactive ASAR (white clock) on each homolog. The red and green clouds surrounding the
839 chromosomes represent "ASAR" RNA expressed from the different active "ASARs" on
840 each homolog.
841

Table 1. Coordinated Asynchronous Replication Timing by ReTiSH.

PBLs					
Locus 1	**Locus 2**		**cis (%)**	**trans (%)**	**P value**
ASAR6-141	ASAR6		75	25	<1 x 10-3

HTD114					
Locus 1	**Locus 2**		**cis (%)**	**trans (%)**	**P value**
ASAR6-141	CHR6 cen*		77	23	<1 x 10-4
ASAR6	CHR6 cen*		88	12	<1 x 10-6
ASAR6-141	ASAR6		76	24	<1 x 10-4

CHR6 cen* indicates the large centromere on one of the chromosome 6s in HTD114 cells.

842

843

Supporting Information

S1 Fig. Delayed replication of chromosome 6 following disruption of *ASAR6*. **(**A and B) A representative mitotic spread from BrdU (green) treated cells containing a deletion of the expressed allele of *ASAR6 [26]*. Mitotic cells were subjected to DNA FISH using a chromosome 6 centromeric probe (red). The larger centromere resides on the chromosome 6 with the expressed *ASAR6* allele (CHR6A). Bar is 10 uM. (C) The two chromosome 6s were extracted from a and b and aligned to show the BrdU incorporation and centromeric signals. (D) Pixel intensity profiles of BrdU incorporation and DAPI staining along the (6A) and (6B) chromosomes from panel C. The long (q) and short (p) arms of chromosome 6 are indicated. Bar is 2 uM. (E) BrdU quantification along 6A and 6B from panel D. (F) The ratio of DNA synthesis into the two chromosome 6s was calculated by dividing the BrdU incorporation in 6B by the incorporation in 6A.

S2 Fig. Expression and asynchronous replication of *ASAR6*. UCSC Genome Browser view of the *ASAR6* RNA-seq data, showing the reads from the plus and minus strand in separate tracks, from HTD114 nuclear ribo-minus RNA. We previously mapped the ~1.2 mb asynchronous replication domain associated with *ASAR6* as indicated (see [20]). We note that the RNA-seq reads associated with *ASAR6* fulfill all of the characteristics described for vlincRNAs (see [23, 24]). Two additional vlincRNAs (253 and 254) also map to the asynchronous replication domain. Also shown is the Repeat Masker Track.

S1 Table. Repetitive elements within vlinc273 (ASAR6-141). The chromosome position, orientation, size and total bases occupied by LINE, Alu, and other repeats, from RepeatMasker are shown.

S2 Table. DNA oligonucleotides used for sgRNAs and PCR primers. The DNA sequence of the oligonucleotides and the position on chromosome 6 of the oligonucleotides used in for sgRNAs and PCR primers used to screen for deletions. Also shown are the heterozygous SNPs within the PCR products used to determine which allele was expressed and/or deleted following CRISPR/Cas9 expression.

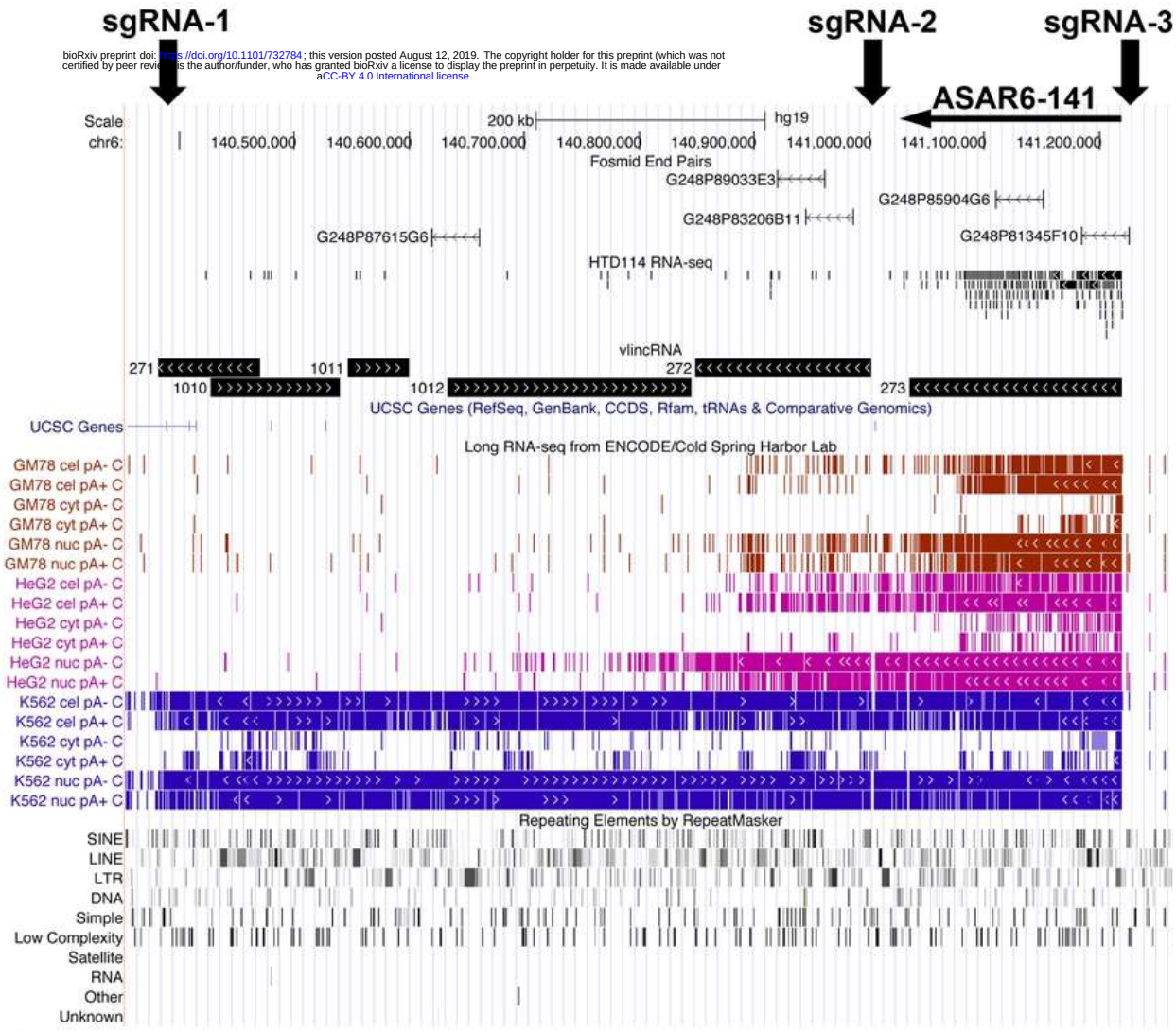


Figure 1

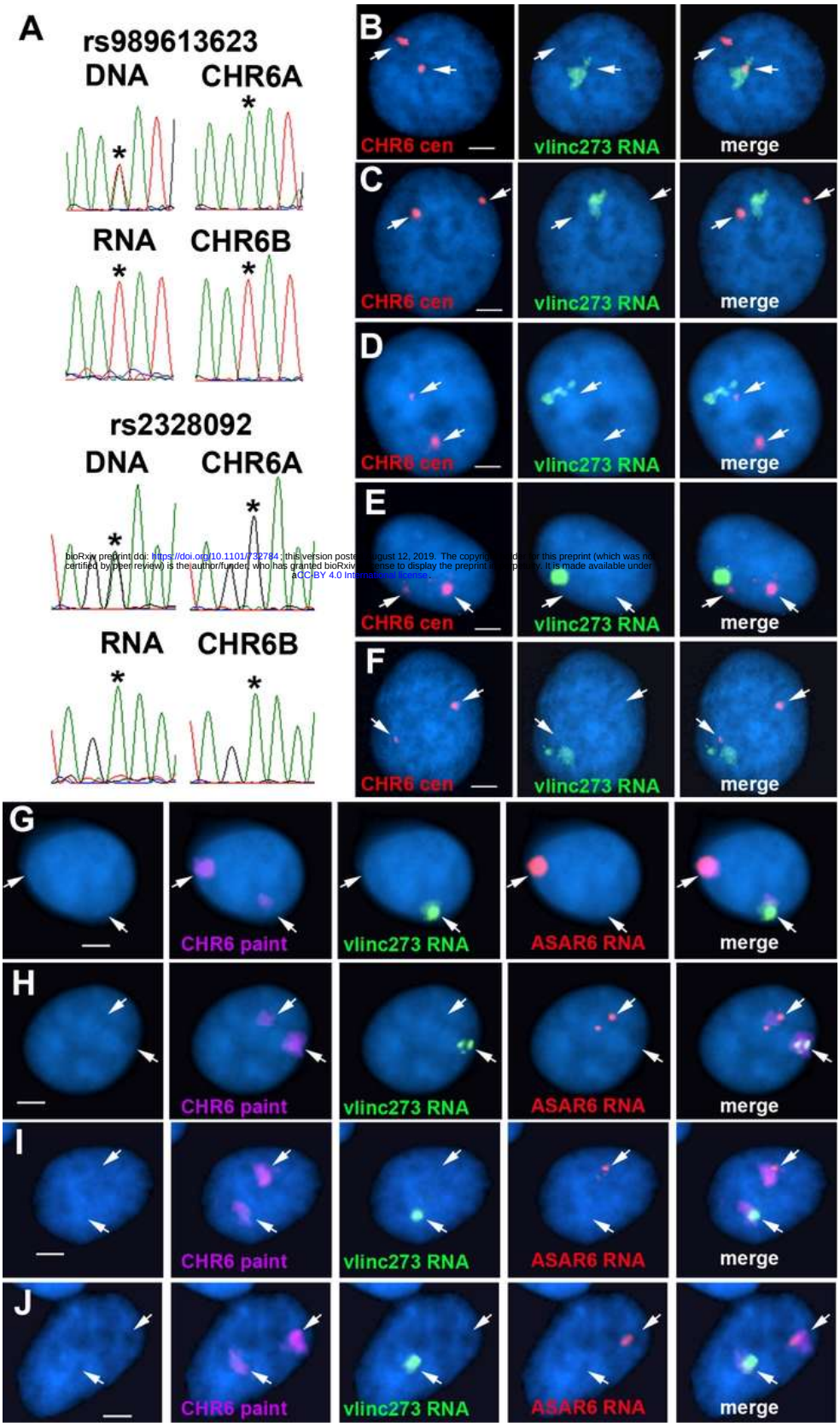


Figure 2

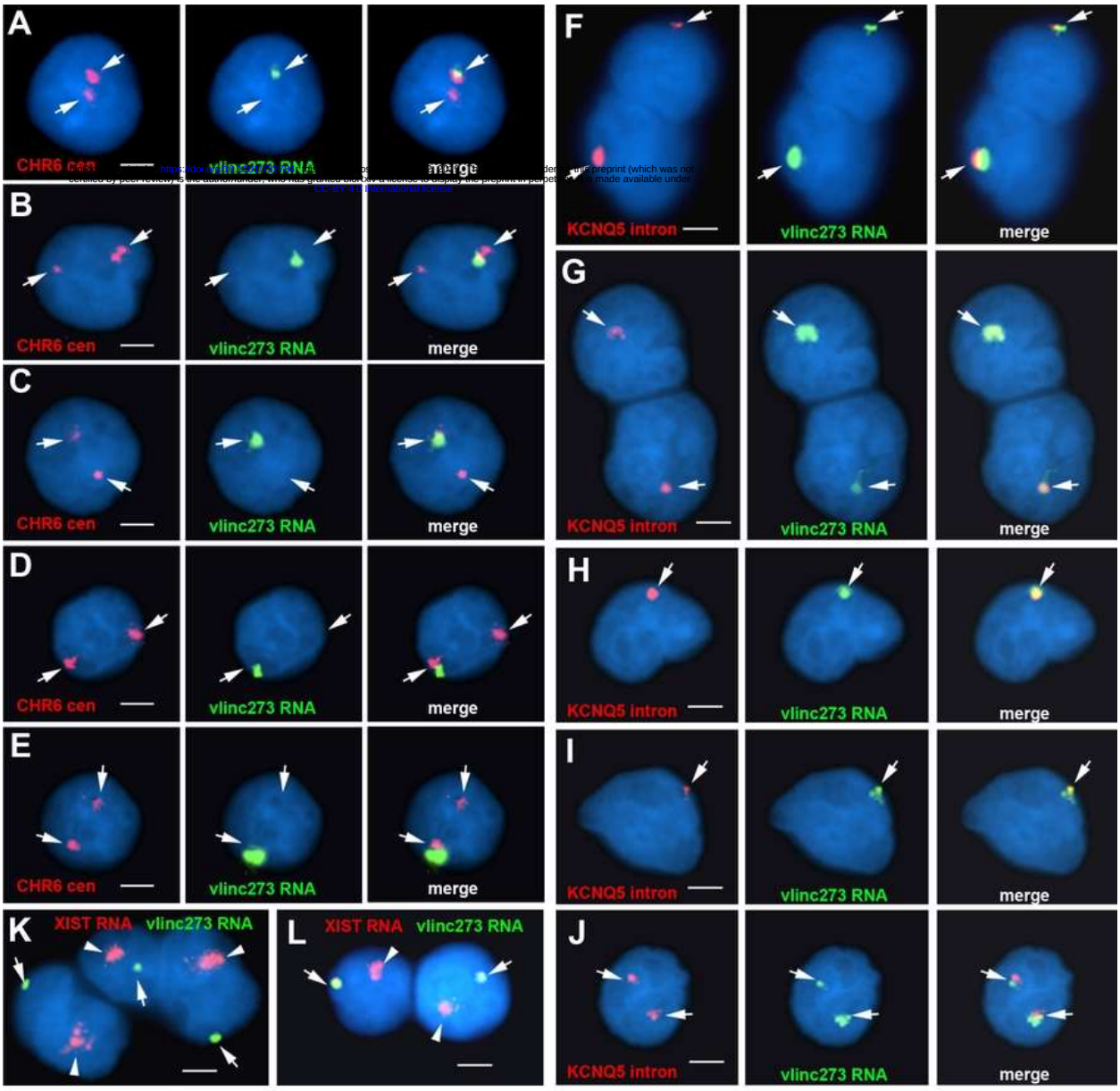

A
CHR6 cen
vlinc273 RNA
merge
B
CHR6 cen
vlinc273 RNA
merge
C
CHR6 cen
vlinc273 RNA
merge
D
CHR6 cen
vlinc273 RNA
merge
E
CHR6 cen
vlinc273 RNA
merge
F
KCNQ5 intron
vlinc273 RNA
merge
G
KCNQ5 intron
vlinc273 RNA
merge
H
KCNQ5 intron
vlinc273 RNA
merge
I
KCNQ5 intron
vlinc273 RNA
merge
J
KCNQ5 intron
vlinc273 RNA
merge
K
XIST RNA
vlinc273 RNA
L
XIST RNA
vlinc273 RNA

Figure 3

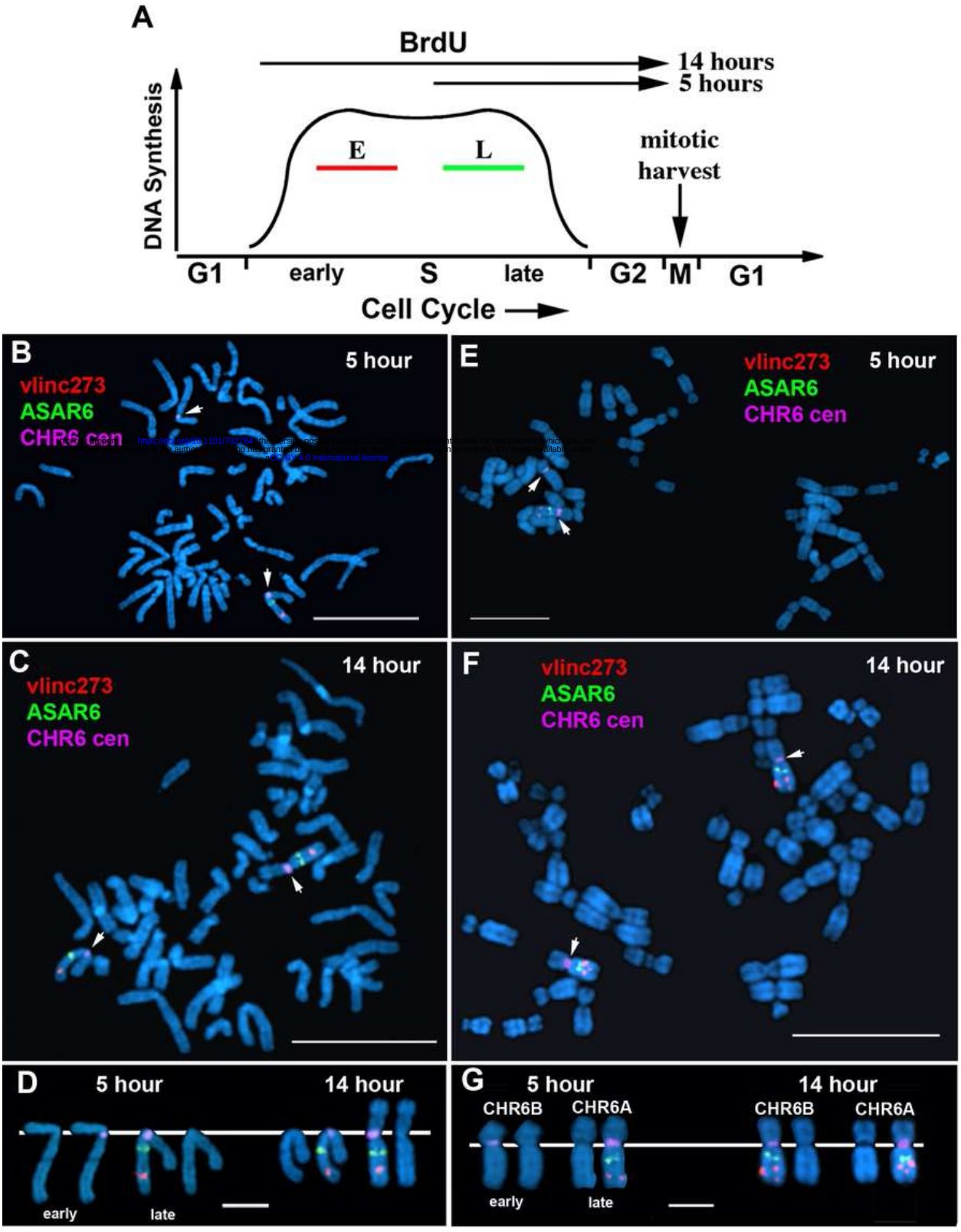
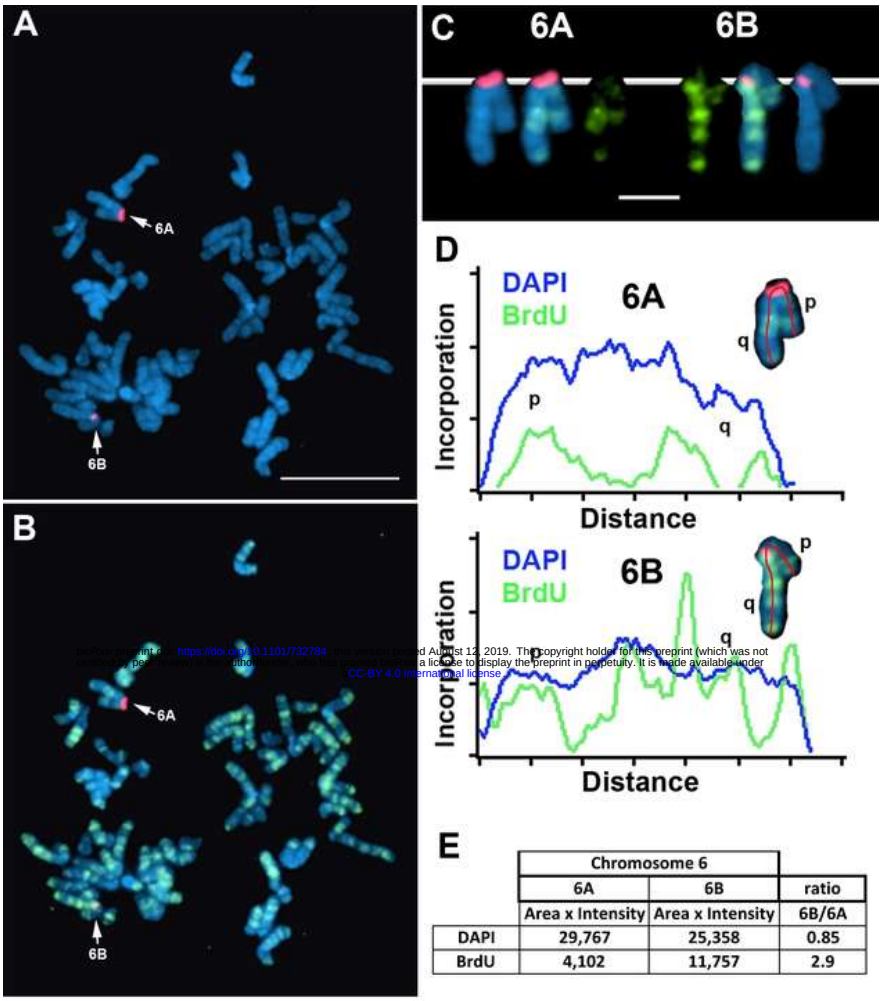

A
BrdU
14 hours
5 hours
DNA Synthesis
E
L
mitotic
harvest
G1
early
S
late
G2
M
G1
Cell Cycle
B
5 hour
vlinc273
ASAR6
CHR6 cen
E
vlinc273
5 hour
ASAR6
CHR6 cen
C
14 hour
vlinc273
ASAR6
CHR6 cen
F
vlinc273
14 hour
ASAR6
CHR6 cen
D
5 hour
14 hour
early
late
G
5 hour
CHR6B
CHR6A
14 hour
CHR6B
CHR6A
early
late

Figure 4



E

	Chromosome 6		
	6A	6B	ratio
	Area x Intensity	Area x Intensity	6B/6A
DAPI	29,767	25,358	0.85
BrdU	4,102	11,757	2.9

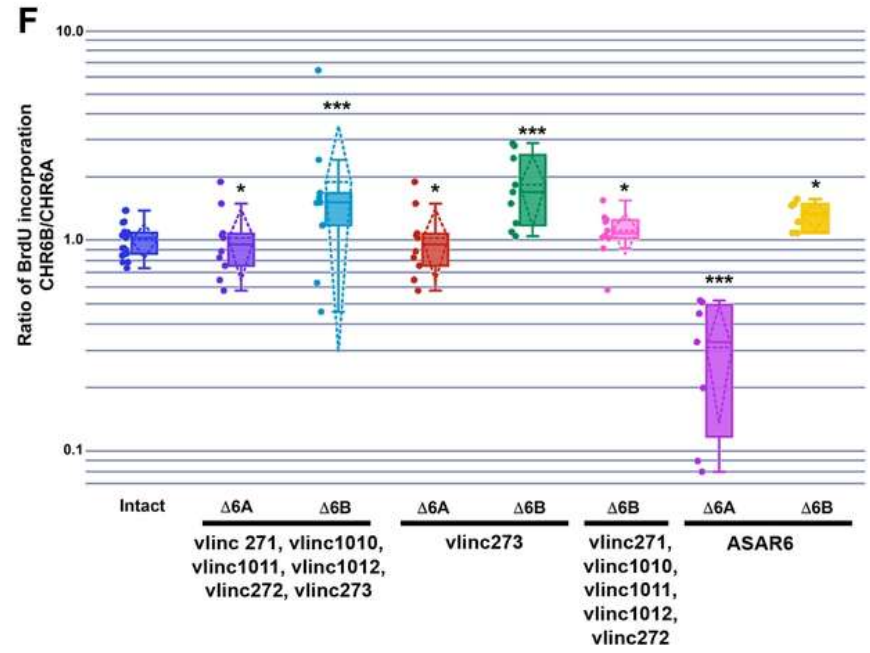


Figure 5

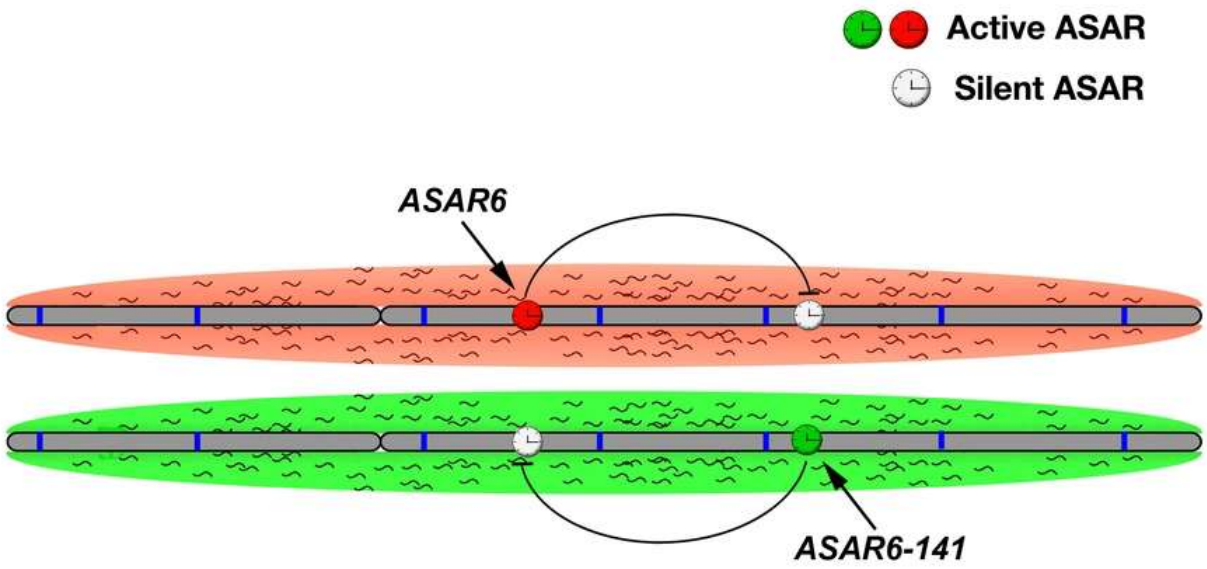


Figure 6

Dissertation Discussion and Future Work

Ovarian Teratoma

The genetic study of ovarian immature teratoma described above has yielded several main findings: 1) immature teratoma genomes contain extremely low levels of SNVs, but contain widespread CN-LOH as a result of meiotic aberrations of germ cells, 2) genomes derived from cells in histologically distinct tumor regions do not contain exclusive mutations, indicative of a highly clonal tumor that has not undergone heterogeneous evolution of subclones, and 3) bilateral ovarian teratomas are genetically independent tumors, and glial cell deposits in the peritoneum (gliomatosis peritoneii) are genetically derived from ovarian primary tumors, indicative of dissemination from the ovary to the peritoneum. I will now speculate at length about the potential significances of each finding, and suggest follow up experiments for the near and distant future.

It is not unexpected that ovarian immature teratoma genomes contain very low levels of SNVs, as teratomas are often diagnosed while individuals are in adolescence and somatic mutation level is thought to be a function of age. This however presents a significant difference compared to ovarian dysgerminoma and analogous male germ cell tumors, where recurrent mutations in the KIT and KRAS genes are found. The KIT tyrosine kinase is known for its role in germ cell migration during embryonic development, thus one plausible model of KIT driven germ cell tumor is that a rare sporadic mutation in KIT leads to abnormal germ cell migration and exposure of the developing germ cell to an environment that somehow increases the risk of a failure of meiosis. Since ovarian immature teratoma lack these KIT mutations and KRAS amplifications, it suggests that an epigenetic process, environmental exposure, or mere random chance could be responsible for failure of meiosis and the resulting CN-LOH that is assumed to be sufficient to drive tumorigenesis. Strengthening this argument is the observation that boys with cryptorchidism, delayed descent of the testicles, and women with disorders of sex development, conditions that both presumably expose germ cells to foreign and potentially harmful environments, put individuals at significantly higher risk of germ cell tumors. This hypothesis could be tested by observing the behaviors and conditions of pregnant women, and looking for any associations with producing

children that suffer from ovarian teratoma. However, due to the 15-25 year delay in ovarian teratoma development from birth, this would only be practical if detailed data for large numbers of pregnant women was obtained and stored long term, the data was accessible to the research community, and the cancer diagnosis in the offspring was connected to the parents data. Alternatively, mouse models of teratoma lacking KIT and KRAS mutations could be utilized to discover factors driving meiotic failure. Finding 2) describes a tumor that contains extremely heterogeneous cellular differentiation, as observed in many other tumor types, but remarkably homogeneous genomes. This morphologically heterogeneous but genetically homogeneous tumor could provide a reasonably well controlled environment to study epigenetic tumor evolution, a phenomenon that is under intense investigation in the context of tumor adaptation to therapy. The study of DNA methylation changes between morphologically distinct tumor regions could yield novel information about epigenetic tumor evolution and is worthy of further study. Although relatively expensive, whole genome bisulfite sequencing would give insight as to the DNA methylation status of morphologically distinct tumor regions and the epigenetic evolution of immature teratoma. This also poses another question as to whether the cellular differentiation composition of ovarian teratomas is associated with clinical behavior and tumor aggressiveness, although this would require a large cohort of well annotated tumors linked to long term clinical data. Lastly, finding 3) demonstrates that teratoma cells are capable of escaping the primary tumor environment, travelling to the peritoneum, and differentiating somehow into glial cells. An interesting follow up question is why and how a germ cell differentiates into glia, a major component of the brain, in the peritoneum, as well as what features of glial cells are well adapted or modified to enable neoplastic growth in the peritoneum. Because gliomatosis peritoneii is genetically derived from primary teratoma, clinicians can treat it as a metastatic lesion instead of considering it an entirely separate clinical entity. We have also found that bilateral tumors are genetically independent and are the result of two unrelated tumors. Thus, bilateral teratoma can be considered as two separate clinical entities, and treatment needs to be designed accordingly.

ASAR lncRNA

The discovery of ASAR6-141 has led us to the hypothesis that all chromosomes may contain multiple ASAR lncRNAs necessary for replication timing and stability. Although not described in detail in this dissertation, we have tested the hypothesis that all chromosomes contain ASARs by utilizing the uncommon qualities of ASAR RNAs including very long size (>50kb), random monoallelic expression of nuclear RNA, and differential allelic replication timing. The model system we have chosen is a haplotype resolved lymphoblastoid cell line, where we were able to detect and isolate six distinct subclonal populations based on immunoglobulin gene rearrangements. The six subclones are virtually genetically identical besides at somatic rearrangement sites at immunoglobulin genes, producing a system that allows us to confidently study epigenetic differences, including replication timing and monoallelic expression, between cell populations that are derived from the same individual but may belong to distinct cell lineages. Each subclone has the same X chromosome inactivated, but belongs to a different B-cell derived lineage, as shown by the presence of unique immunoglobulin rearrangements in each sublcone. We utilized RNA-seq and Repli-seq to study the differential allelic expression landscape of very long non-coding RNAs, and assessed the genomic context of replication timing around these expressed loci. Although the work is currently unfinished, unpublished data has demonstrated that all chromosomes demonstrate differential expression of very long non-coding RNA, with the same nuclear localization pattern as observed in the bona fide ASAR genes. We have observed one potential ASAR gene every ~20mb throughout the genome. We have also begun to validate the predicted ASARs using DNA/RNA FISH visualization, and gold standard BrdU incorporation assays to observe changes in replication timing after creating allele specific genomic deletions. We have found that the predicted loci do in fact have the ASAR expression pattern and are required for normal replication timing. I will now extrapolate this finding and speculate about the full function of ASAR genes. ASARs are required for normal replication timing and chromosome stability. What processes of DNA replication could require a very long non-coding RNA to function? One idea is that ASARs are necessary for the 3-Dimesional localization of chromosomes which is somehow crucial to replication. It is known that early replication regions tend to

be located more centrally within the nucleus, and late replicating regions on the nuclear periphery, thus ASARs could somehow assist in this spatial regulation, however it is not clear why subnuclear genome localization is important to genome stability. lncRNAs are known to be associated with nuclear matrix proteins, and can be involved in ribonucleoprotein superstructures, and therefore could act as a scaffold of genome organization, but more work into the biochemical basis of lncRNA genome contact is needed. Another possibility is that ASARs regulate allele specific gene regulation of protein-coding genes that are critical to DNA replication. This idea has some precedent: several other genes require allele-specific regulation for critical biological processes as mentioned above such as at the Prader-Willi/Angelman locus, X inactivation center, imprinted lncRNA loci, and immune/olfactory loci. If true, the next question would be to identify which protein coding genes require allele-specific expression for proper DNA replication. Also, what benefit is achieved through the evolution of a more complex regulation scheme for a process like DNA replication that is fundamental to all life and occurs at large scale throughout the lifetime of organisms? Now consider the finding that ASARs are required for chromosome stability. Is this due to ASAR essentiality in normal replication timing, or a process independent of replication? Hypotheses that we can now put forth are that ASARs assist with chromosome structure, and/or regulate allele-specific expression of an array of protein coding genes that are then responsible for critical cellular processes. Lastly, let us consider the impact of ASAR research on human health. Because instability is a nearly omnipresent feature of cancer cells, and the loss of ASAR genes leads to chromosome stability, we hypothesize that loss of ASAR genes may be important in tumorigenesis. We can test this hypothesis by analyzing a variety of tumor cells for ASAR expression, however this may be more successfully achieved after a thorough cataloging of ASAR genes is done in many normal tissues. Genome instability is currently accepted as a means by which genomes rapidly gain or lose oncogenes and tumor suppressor genes, allowing for rapid acquisition of tumorigenicity and/or adaptation to therapeutic pressure, thus loss of ASAR expression could be an important oncogenic step at any stage of cancer development. Because ASARs are located in the non-coding portion of the genome, knowledge of ASAR mutations in tumor cells could provide insight that assessment of the coding genome does not provide. Additionally, because

ASARs are biomarkers of chromosome stability, they could be utilized for diagnostic and prognostic purposes.

Appendix

Michael has co-authored the following papers during the graduate school training period.

Fei, S. S., et al. (2016). "Patient-specific factors influence somatic variation patterns in von Hippel-Lindau disease renal tumours." Nat Commun **7**: 11588.

Cancer development is presumed to be an evolutionary process that is influenced by genetic background and environment. In laboratory animals, genetics and environment are variables that can largely be held constant. In humans, it is possible to compare independent tumours that have developed in the same patient, effectively constraining genetic and environmental variation and leaving only stochastic processes. Patients affected with von Hippel-Lindau disease are at risk of developing multiple independent clear cell renal carcinomas. Here we perform whole-genome sequencing on 40 tumours from six von Hippel-Lindau patients. We confirm that the tumours are clonally independent, having distinct somatic single-nucleotide variants. Although tumours from the same patient show many differences, within-patient patterns are discernible. Single-nucleotide substitution type rates are significantly different between patients and show biases in trinucleotide mutation context. We also observe biases in chromosome copy number aberrations. These results show that genetic background and/or environment can influence the types of mutations that occur.

Co-author contribution: Michael performed a mutation signature analysis of VHL tumors and observed patient specific mutation signatures that were not shared between patients.

Wang, Y., et al. (2018). "Long-Term Correction of Diabetes in Mice by In Vivo Reprogramming of Pancreatic Ducts." Mol Ther **26**(5): 1327-1342.

Direct lineage reprogramming can convert readily available cells in the body into desired cell types for cell replacement therapy. This is usually achieved through forced activation or

repression of lineage-defining factors or pathways. In particular, reprogramming toward the pancreatic β cell fate has been of great interest in the search for new diabetes therapies. It has been suggested that cells from various endodermal lineages can be converted to β-like cells. However, it is unclear how closely induced cells resemble endogenous pancreatic β cells and whether different cell types have the same reprogramming potential. Here, we report in vivo reprogramming of pancreatic ductal cells through intra-ductal delivery of an adenoviral vector expressing the transcription factors Pdx1, Neurog3, and Mafa. Induced β-like cells are mono-hormonal, express genes essential for β cell function, and correct hyperglycemia in both chemically and genetically induced diabetes models. Compared with intrahepatic ducts and hepatocytes treated with the same vector, pancreatic ducts demonstrated more rapid activation of β cell transcripts and repression of donor cell markers. This approach could be readily adapted to humans through a commonly performed procedure, endoscopic retrograde cholangiopancreatography (ERCP), and provides potential for cell replacement therapy in type 1 diabetes patients.

Co-author contribution: Michael drove the single-cell RNA-seq analysis of reprogrammed pancreatic ductal cells, showing that reprogrammed cells were similar to native beta cells, but not indistinguishable.

Boniface, C., et al. (2021). "The Feasibility of Patient-Specific Circulating Tumor DNA Monitoring throughout Multi-Modality Therapy for Locally Advanced Esophageal and Rectal Cancer: A Potential Biomarker for Early Detection of Subclinical Disease." Diagnostics (Basel) **11**(1).

As non-operative management (NOM) of esophageal and rectal cancer is becoming more prevalent, blood-biomarkers such as circulating tumor DNA (ctDNA) may provide clinical

information in addition to endoscopy and imaging to aid in treatment decisions following chemotherapy and radiation therapy. In this feasibility study, we prospectively collected plasma samples from locally advanced esophageal (n = 3) and rectal cancer (n = 2) patients undergoing multimodal neoadjuvant therapy to assess the feasibility of serial ctDNA monitoring throughout neoadjuvant therapy. Using the Dual-Index Degenerate Adaptor-Sequencing (DIDA-Seq) error-correction method, we serially interrogated plasma cell-free DNA at 28-41 tumor-specific genomic loci throughout therapy and in surveillance with an average limit of detection of 0.016% mutant allele frequency. In both rectal cancer patients, ctDNA levels were persistently elevated following total neoadjuvant therapy with eventual detection of clinical recurrence prior to salvage surgery. Among the esophageal cancer patients, ctDNA levels closely correlated with tumor burden throughout and following neoadjuvant therapy, which was associated with a pathologic complete response in one patient. In this feasibility study, patient- and tumor-specific ctDNA levels correlated with clinical outcomes throughout multi-modality therapy suggesting that serial monitoring of patient ctDNA has the potential to serve as a highly sensitive and specific biomarker to risk-stratify esophageal and rectal cancer patients eligible for NOM. Further prospective investigation is warranted.

Co-author contribution: Michael designed a simple Bayesian statistical test to detect the presence of mutations in ultra deep sequencing data from cell-free DNA.

FINISH