CANCER DRIVER-PASSENGER DISTINCTION VIA SPORADIC HUMAN AND DOG COLORECTAL CANCER COMPARISON

by

JIE TANG

(Under the Direction of Shaying Zhao)

ABSTRACT

Human colorectal cancer (CRC) is one of the better-understood systems for studying the genetics of cancer initiation and progression. To develop a cross-species comparison strategy for identifying CRC causative gene alterations, we performed array comparative genomic hybridization (aCGH) to investigate CNAs in 10 sporadic canine CRCs. The results revealed for the first time a strong degree of genetic homology between canine and human CRCs. First, when mapping the identified CNAs onto syntenic regions of the human genome, we noted that the canine orthologs of genes participating in known human CRC pathways were recurrently disrupted, indicating that these pathways might be altered in the canine CRCs as well. Second, we observed a significant overlapping of CNAs between human and canine tumors, and tumors from the two species were clustered according to the tumor subtypes but not the species. Significantly, we found that species-specific CNAs localize to evolutionarily unstable regions. These findings indicate that CNAs recurrent between human and dog CRCs may have a higher probability of being cancer-causative, compared with CNAs found in one species only.

As an expansion, we used this novel human-dog comparison strategy to address a central aim of cancer research: cancer driver–passenger distinction. We compared the CNAs of our 10 canine CRCs to 29 human CRCs. This led to the identification of 73 driver candidate genes (DCGs) altered in both species, and 38 passenger candidate genes (PCGs) altered in humans only. We noted that DCGs significantly differ from PCGs in every analysis conducted. Importantly, while PCGs are not enriched in any specific functions, DCGs possess significantly enhanced functionality in establishing and maintaining epithelial cell polarity. Meanwhile, many DCGs also participate in processes that regulate cell proliferation and death. These observations indicate that, in sporadic CRCs, cell polarity genes not only play a critical role in preventing cancer cell invasion and spreading, but also likely serve as tumor suppressors by modulating cell growth. This pilot study validates our novel strategy, expansion of which would make more driver-passenger distinctions, and address key questions regarding the relationship between cancer pathogenesis and epithelial cell polarity control

INDEX WORDS: Canine Colorectal Cancer; Copy Number Alterations; Cross-Species Comparison; Cancer Driver Passenger Distinction; Epithelial Cell Polarity; Potential Tumor Suppressors;

CANCER DRIVER-PASSENGER DISTINCTION VIA SPORADIC HUMAN AND DOG COLORECTAL CANCER COMPARISON

by

JIE TANG

BS, Northeast Forestry University, China, 2003

MS, University of Science and Technology of China, China, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

CANCER DRIVER-PASSENGER DISTINCTION VIA SPORADIC HUMAN AND DOG COLORECTAL CANCER COMPARISON

by

JIE TANG

Major Professor: Shaying Zhao
Committee: J. David Puett
Paul Schliekelman
Heinz-Bernd Schüttler

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2012

DEDICATION

This dissertation is dedicated to my love for science and research.

ACKNOWLEDGEMENTS

I would like to thank my advisor Shaying Zhao for her support and guidance over the years. She has been instrumental in my scientific training and taught me how to conduct scientific investigations. I would also like to thank the members of my advisory committee: Drs. J. David Puett, Paul Schliekelman and Heinz-Bernd Schüttler, for their mentorship and guidance during my graduate training. I would like to thank all past and present members of the Zhao lab for their friendship. I would like to thank my wife Wenjuan for standing beside me throughout my Ph.D. studies.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION TO CANCER

What is cancer?

Cancer, according to Encyclopedia Britannica, is a generic term describing more than 100 distinct diseases characterized by the uncontrolled growth of abnormal cells in the body, which is known medically as malignant tumor and neoplasm. Basically, it is a condition of aberrant genetic programming (Wong et al. 2011), where changes in the genomic sequences alter the function and expression of genes that regulate essential biological processes. Cancer develops when the normal regulatory mechanisms controlling cell growth, differentiation and death mechanisms are disrupted.

Cancer begins in cells and thus is named for the organ/tissue and type of the cell where it starts. Generally speaking, cancer can be classified into five major categories. Each of them is defined in the National Cancer Institute's (NCI) Dictionary of Cancer Terms as follows:

- Carcinoma- cancer that begins in the skin or in tissues that line or cover internal organs. It is the most common type of cancer, which accounts for 80-90% of all cancer diseases.

- Sarcoma- cancer that begins in bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue.
- Leukemia- cancer that starts in blood-forming tissue such as the bone marrow and causes large numbers of abnormal blood cells to be produced and enter the blood.
- Lymphoma and myeloma- cancers that begin in the cells of the immune system, such as the glands, nodes of the lymphatic system or the plasma cells of bone marrow.
- Central nervous system cancers- cancers that begin in the tissues of the brain and spinal cord.

In addition, some cancers are of mixed types (e.g., adenoneuroendocrine carcinomas).

History of cancer study

Although it is reported that up to 30% of cancer deaths are associated with the dietary and behavioral risks, including obesity, smoking and physical inactivity (Bray et al. 2012), it is obvious that cancer is not just a disease of the modern world. According to American Cancer Society reports (http://www.cancer.org), the oldest known specimen of a human cancer could be dated back to 1900-1600 BC, which is considered to be a head and neck cancer from a woman. The oldest documentation of a cancer case is found to be from ancient Egypt and written on papyri in approximately 3000 BC describing a breast cancer (Hajdu 2011).

The term of "cancer" is believed to have been named by the Greek physician Hippocrates (460-370 BC). He used "*carcinos"* (or *karkinos* in some other reports),

which in Greek refers to a crab, for the first time to describe “non-ulcer forming and ulcer-forming tumors” that looked like a crab to him (Diamandopoulos 1996). This word was later translated by the Roman physician, Celsus (25 BC-50 AD) into the Latin word cancer (refers to a crab), from which cancer in English was derived.

Although the potential causes of cancer have been recorded for more than 2000 years, cancer was generally treated by blood-letting, diet, and/or laxatives, which were suggested by Hippocrates and his disciples, and considered as incurable until the 19th century (Greenwald and Dunn 2009). In 1855, the German doctor Rudolf Virchow (1821-1902) recognized that cancer arises from embryonic-like cells and is caused by chronic irritations (Wagner 1999).

The genetic mechanism of cancer was determined by the German zoologist Theodor Boveri (1862-1915). In 1902, he reasoned that “malignant tumors might be the consequence of a certain abnormal chromosome constitution, which in some circumstances can be generated by multipolar mitoses” (Boveri 2008), and concluded that cancer is a genetic disease of somatic cells, resulting from the poisons, radiation, physical insults, pathogens, chronic inflammation and tissue repair. Also, he suggested the existence of oncogenes, tumor suppressors and cell cycle check points and predicted the sensitivity of cancer to radiotherapy (Satzinger 2008).

Cancer stem cells (CSC) were firstly identified by Jacob Furth and Morton Kahn in 1937 (DeVita and Chu 2008). They demonstrated that mouse leukemia could be

transplanted from a single cell. Until 1960s, people confirmed that only a small fraction of tumor cells were able to proliferate *in vivo* (Bruce and Van Der Gaag 1963) and thus introduced the concept of CSC.

In 1970s, David Comings realized cancer could arise through the inactivation of tumor suppressors (Comings 1973), and the first tumor suppressor *RB1* was determined ten years later (Cavenee et al. 1983). At the same time, the first oncogene v-*src* was discovered in a chicken retrovirus (Martin 1970) and confirmed to be the activated proto-oncogene c-*src* in human (Spector et al. 1978).

Studies in the 1980s demonstrated that epigenetic abnormalities in cancer could underlie oncogene activation as well as the silencing of tumor suppressors (Feinberg and Vogelstein 1983; Greger et al. 1989). Another milestone in cancer studies is the understanding of *TP53* and *RB1* roles in cell-cycle and DNA-damage checkpoints (Buchkovich et al. 1989; DeCaprio et al. 1989; Kastan et al. 1991), which have been the focus of cancer research for the past decade.

In the late 1990s, the microarray technology was sophisticated enough to enable distinguishing between cancer types for better clinical outcome predictions (Golub et al. 1999). Since then, the cancer profiling analysis by high throughput array and next generation sequencing has become an essential and routine tool in cancer research, diagnostics as well as the targeted therapy.

Cancer prevalence

It is reported that cancer is a leading cause of death worldwide, accounting for 7.8 million cases, and the second leading cause of death in the United States, accounting for 0.6 million cases in 2008 (Ferlay et al. 2010; Howlader et al. 2010). The incidence in both sexes continue to increase, with 12.7 million new cancer cases worldwide in 2008, and up to 13.1 million deaths are expected from cancer in 2030 (Ferlay et al. 2010). It is estimated that about 12 million cancer cases have been diagnosed in the Unites States as of January 2008. Among them, 77 percent of cancers occur in people above the age of 55 (Howlader et al. 2010).

MUTATIONS IN CANCER

Cancer is a disease of the genome

Nowadays, it is widely accepted that cancer is a disease of the genome (Stratton et al. 2009). A typical cancer genome may harbor well over 3,000 sequence mutations and hundreds of chromosomal aberrations (Macconaill and Garraway 2010). Among such events, driver mutations are those that confer a growth/survival advantage and are positively selected during cancer development and progression. This is in contrast to passenger events, which do not contribute to the development of cancer but are carried along during clonal expansion (Wong et al. 2011). As of February 2012, a total of 474 genes have been documented as having known cancer driver mutations in Cancer Gene Census (CGC) database http://www.sanger.ac.uk/genetics/CGP/Census/). Among these known driver events, 90% of them are somatic mutations while only 20% are germline changes, including the well known *APC*, *BRCA1* and *BRCA2* mutations. Based on the

CGC data, it is believed that most of somatic mutations (~90%) are oncogenetic and about 10% function as tumor suppressors.

The passenger mutations, on the other hand, are generally determined by mutation frequency in the genome, which is associated with the genetic defects in the DNA repair processes, and can differ very significantly across different tumors. For example, it is reported that the minimum estimate of passenger mutation rate for colon and breast cancers are 0.55 and 0.33 per Mb, which account for about 1,650 and 1,000 mutations, respectively (Stephens et al. 2005; Wood et al. 2007).

In most cases, passenger alterations vastly outnumber these driver events (Stratton et al. 2009). For example, only 18 and 15 driver genes have been documented in colon and breast cancers in CGC and for a single tumor; it is estimated that less than 15 mutations are drivers (Wood et al. 2007) while more than 1,650 and 1,000 passenger mutations are expected to be found (Stephens et al. 2005; Wood et al. 2007). Indeed, in the most recent Release 57 of Catalogue of Somatic Mutations in Cancer database (COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic/), as many as 75,109 unique somatic mutations have been recorded across 15,236 genes while only 474 of them are known to be drivers (Futreal et al. 2004; Forbes et al. 2008).

Hence, a central aim and a big challenge in cancer studies are to distinguish drivers from passengers. Many large-scale projects have been initiated in the past few years and intense efforts have been made to address this key issue (Greenman et al. 2007;

Ding et al. 2008; Jones et al. 2008). These studies have identified multiple driver genes that have been discovered through experimental efforts over the past 30 years as well as many new candidates. Most efforts on identification of candidate driver genes are based on the assumption as follows.

- The common highly mutated genes among multiple cases within one or multiple tumor types are likely the most important drivers (Sjoblom et al. 2006; Greenman et al. 2007).
- Non synonymous mutations clustered in protein-coding or regulatory sequences tend to be drivers, especially those located at loci which are known to be cancer relevant such as protein kinase coding regions (Stratton et al. 2009).

It is expected that improved techniques like genome sequencing will uncover many more somatic mutations distributed throughout the genome with yet unknown potential functional effects.

Another promising strategy is the comparative oncogenomics method, which studies cancer genomes and transcriptomes by the cross-species comparison to characterize genes associated with cancer (Peeper and Berns 2006). Comparative oncogenomics has been demonstrated to be a powerful approach in identification of driver genes in oncogenesis and metastasis in human tumors (Kim et al. 2006; Zender et al. 2006). In comparative cancer research, the mouse is the most commonly used model organism. Mice have a shorter life span, are small in size, commercially available and cheaper to raise compared with other mammals. Furthermore, the mouse is genetically best characterized of all mammals and the most experimentally tractable mammalian

systems (Sharpless and Depinho 2006). Nowadays, next generation gene targeting in mouse models offers the opportunity to manipulate the expression of genes within a few selected cells (Gondo et al. 2009). However, there are some notable limitations in mouse models in cancer study. One is that tumors in human are naturally occurring and are primarily caused by the accumulation of somatic mutations while tumors in the mouse are genetically engineered (Cheon and Orsulic 2011) and thus the models often involves only one or a few genes and fails to recapitulate the heterogeneous nature of cancer (Rowell et al. 2011). Therefore mouse models are usually unable to reveal gene networks and interactions that are responsible for human cancers (Karlsson and Lindblad-Toh 2008; Rowell et al. 2011).

Copy number alterations in cancer

In cancer, somatic mutations usually take the form of single base substitutions, as well as small insertions/deletions and structural variations (Klopocki and Mundlos 2011). Structural variations often refer to the changes of genomic DNA greater than 50bp in size, which include the unbalanced copy number changes and the balanced changes (i.e., inversions, reciprocal translocations). Copy number changes or copy number alterations (CNAs), by definition, are alterations of the DNA ranging from 50bp to the whole chromosome arm, which result in the cell having an abnormal number of copies of one or more sections of the DNA (Stankiewicz and Lupski 2010). It can disrupt the biological balance of normal ploidy by deletions (fewer than the normal number) and amplifications (more than the normal number) on certain chromosomes. As one of the main mechanisms

causing cancer related alterations, CNAs are frequently observed in epithelial cancers (e.g., SKY/M-FISH and CGH Database, http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi).

CNAs can cause a cancer by several molecular mechanisms. One of the best known mechanisms is the gene dosage effect , that is, changes in copy number of a gene can alter the amount of RNA and protein produced, therefore leading to significant phenotypic consequences (Cappuzzo et al. 2005). Another commonly recognized mechanism is the position effect, perturbing long-range gene regulations (Kleinjan and van Heyningen 2005). It has been reported that such position effects could be exerted when the breakpoints are located as far as 2–7 Mb away from the causative gene (Merla et al. 2006; Stranger et al. 2007). Copy number deletions also could result in loss of heterozygosity (LOH), which is the loss of one allele in a locus where the other allele was already inactivated or masked. In tumor suppressor loci, heterozygous cells that harbor one normal and one mutated allele still can behave normally. However, LOH lead to the deletion of normal copy, which could predispose the somatic cell to tumor formation (Choate et al. 2010).

Platforms of CNA study

It has been about 20 years (Kallioniemi et al. 1992) that CNAs on metaphase chromosomes have been detected by chromosomal comparative genomic hybridization (CGH), or conventional CGH. Due to its inability of detecting deletions less than 5Mb and amplifications less than 1Mb in length (Klopocki and Mundlos 2011), fluorescence in situ hybridization (FISH), pulsed-field gel electrophoresis (PFGE), and multiplex

ligation-dependent probe amplification (MLPA) techniques were used to identify small copy number deletions until the application of microarray based CGH (aCGH) in late 1990s (Solinas-Toldo et al. 1997).

The older aCGH platforms, consisting of an ordered set of defined bacterial, yeast or P1 artificial chromosome clones (BACs,YACs or PACs) in substitution of the conventional chromosome targets, offer a resolution typically of 100 kb. In recent years, oligo aCGH, based on attaching 25-85 mer oligonucleotide probes as arrays on glass slides, provides a genome-wide copy number profile at resolution as high as 200bp (Urban et al. 2006; Ylstra et al. 2006). More and more new oncogenes and tumor suppressors have been identified by the high density aCGH (Kallioniemi 2008). It has opened a new field of studying the variation of the human genome and its relation to phenotypes and diseases, such as mental retardation (Stankiewicz and Beaudet 2007) and cancer studies (The Cancer Genome Atlas project (TCGA), http://tcga-data.nci.nih.gov/tcga/).

In addition to the array-based methods (i.e., aCGH, single nucleotide polymorphisms (SNP) array), the next generation sequencing (NGS) has further enabled the identification of cancer causative CNAs (Wood et al. 2007; Chiang et al. 2009). NGS technologies have made the cancer sequencing faster, cheaper and more sensitive, and thus have been adopted by many large scale cancer genome projects (e.g., the International Cancer Genome Consortium (ICGC), http://www.icgc.org). It is expected that the NGS is the technology that will prevail in cancer study in the future.

COLORECTAL CANCER

Colorectal cancer (CRC) is one of most well-studied cancer types in humans

CRC is cancer that starts in the large intestine (colon) or the rectum (end of the colon). It is the third most common type (10%) of cancer in the United States (Jemal et al. 2007) and the third leading cause of cancer-related death (8-9%) in the Western world (Altekruse et al. 2009). Considerable progress has been made over the past 20 years in defining the molecular basis of CRC. As a result, CRC is a good model for the study of genetic stages in cancer initiation and progression (Kinzler and Vogelstein 1996, figure 1.1). CRC is a multi-pathway disease and genomic instability is central to its oncogenesis: chromosomal instability (CIN) and microsatellite instability (MSI) (Guastadisegni et al. 2010). MSI is defined by a wide spread alteration of microsatellite regions, usually mono-, di-, or trinucleotide repeats, due to defective DNA mismatch repair (MMR) system. CIN is defined by CNAs of relatively narrow genomic regions (i.e. aneusomy, gains and losses of chromosomal regions). Currently, there are more than 250 CRC genomic studies on more than 3000 samples deposited in a public genome database (GEO, http://www.ncbi.nlm.nih.gov/geo/). Hundreds of human CRCs have been and will be sequenced in the large-scale cancer projects (http://www.icgc.org).

OUR EXPERIMENTAL DESIGN AND INNOVATION

The sporadic canine CRCs are excellent models for studying human CRCs

Instead of traditional mouse models, dog cancer serves as a better model analogous to humans and thus has been applied in our comparative oncogenomics study for several reasons.

- As a companion animal, dogs share the same environment as humans and thus exposed to the same environmental carcinogens. In addition, the dog genome has been sequenced at ~7.5 fold average coverage and a relative accurate reference genome has been available to the public (Lindblad-Toh et al. 2005). On the other hand, the genome of the other pet, cat, was shallowly sequenced (~ 2X coverage) (Pontius et al. 2007), which makes dog be the best available animal model for epidemiological studies of diseases common to human and dog, such as cancer.

- Dogs receive a high level (second only to human) of health care (over $40 billion per year in US, (American Veterinary Medical Association, 2007)) and are kept well into their old age. As many as 45% of them are more than six years old in 2006 (American Veterinary Medical Association, 2007), equivalent to 60–95 years of age in human. Thus, the dog can be the best model for research on genetic and environmental factors to human cancer, which primarily is a disease of aging (Howlader et al. 2010).

- The dog DNA sequences are more similar to humans than are those of the most widely used model organism, the mouse (Lindblad-Toh et al. 2005). About 650 Mb of ancestral sequence is shared between human and dog (not in mice), and thus, many aspects of human biology are presumably more relevant to dogs than mice (Rowell et al. 2011).

- Dog cancers arise spontaneously and are heterogeneous, sharing clinical biological similarities to naturally occurring human cancers (Rowell et al. 2011). It is reported that dog tumors are anatomically and histologically similar to humans and respond similarly to conventional therapies with the same type of cancer (Khanna et al. 2006; Paoloni and Khanna 2008).

- Dog tumor samples are easily accessed. Many efforts have been made to construct a biospecimen repository (Canine Hereditary Cancer Consortium (CHCC), http://www.tgen.org/research/index.cfm?pageid=1382) and a mechanism to share reagents and resources in the canine cancer community has been developed (Canine Comparative Oncology & Genomics Consortium (CCOGC), http://www.ccogc.net/).

- The dog genome is substantially rearranged in comparison to humans with more than 300 inversions and translocations identified (Lindblad-Toh et al. 2005; Ji and Zhao 2008), resulting in thousands of genes that are clustered in the human genome to be dispersedly located in the dog genome. We take advantage of the differences in genomic locations between orthologous genes of the two species to sort driver and passenger candidates (Tang et al. 2012). This fundamentally enables our cross-species comparison strategy to better distinguish drivers from passengers than human cancer-only approaches on cancers with structural variations, to which the vast majority of cancers belong, as described below.

Our comparative oncogenomics strategy sorts genes altered in cancer into driver or passenger candidates

In response to the strong demand for cancer driver identification, we are developing a cross-species comparison strategy that is novel in both concept and approach. One big disadvantage of the previous frequency-based and human-only studies is that the propensity of recurrent mutations does not necessarily imply that recurrent alterations are drivers (Torkamani and Schork 2009; Macconaill and Garraway 2010), especially for those located in evolutionary unstable genomic sites. These are defined as regions enriched with inter-species genomic rearrangement breakpoints (Pevzner and Tesler 2003; Zhao et al. 2004).

Our human-dog comparative strategy hypothesize that driver alteration candidates can be distinguished from passenger candidates by cross-species investigation of the orthologous genomic loci with the same type (and subtype) of cancer. Providing that the two species share similar molecular pathways of oncogenesis, abnormalities that are recurrent between the two species will be considered driver events, whereas those that are found in only one species and located in evolutionary breakpoint sites will be considered passenger events (Figure 1.2).

To test our hypothesis and validate this novel cross-species comparative approach for driver – passenger distinction, we conducted a proof of principle study with CRC. First, we demonstrate canine CRC share similar molecular pathway in carcinogenesis as their human counterparts. Then, we applied the cross-species approach on driver-

passenger distinction. Our studies demonstrate this novel dog-human comparison strategy is more robust than human-only approaches for driver-passenger distinction for cancers with large genomic amplifications/deletions.

FIGURE LEGENDS

Figure 1.1 The human colorectal tumorigenesis model proposed by Vogelstein and colleagues. Colorectal tumorigenesis is initiated by *APC/β-catenin* mutations in WNT pathway. The tumors progresses as the inactivation of other tumor suppressors and activation of oncogenes. The process is accelerated by the development of genomic instability.

Figure 1.2 Our cross-species comparison for driver identification. Abnormalities shared by both species are likely drivers (blue); those uniquely found in one species and located in genomic rearrangement breakpoints are likely passengers (red in the middle panel). An example of driver-passenger distinction on human 5q22.2 via human-mouse comparison is shown in the bottom panel (Ji et al. 2010).

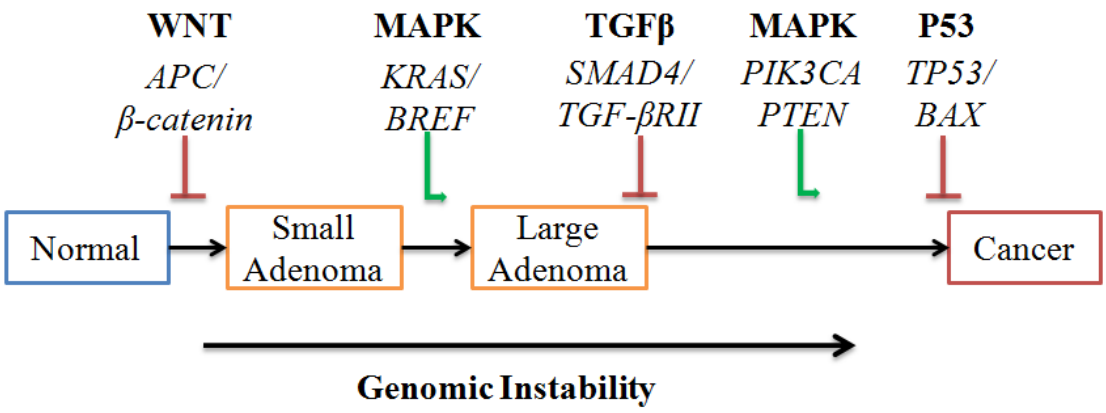

WNT
APC/
β-catenin
MAPK
KRAS/
BREF
TGFβ
SMAD4/
TGF-βRII
MAPK
PIK3CA
PTEN
P53
TP53/
BAX
Normal
Small Adenoma
Large Adenoma
Cancer
Genomic Instability

Figure 1.1

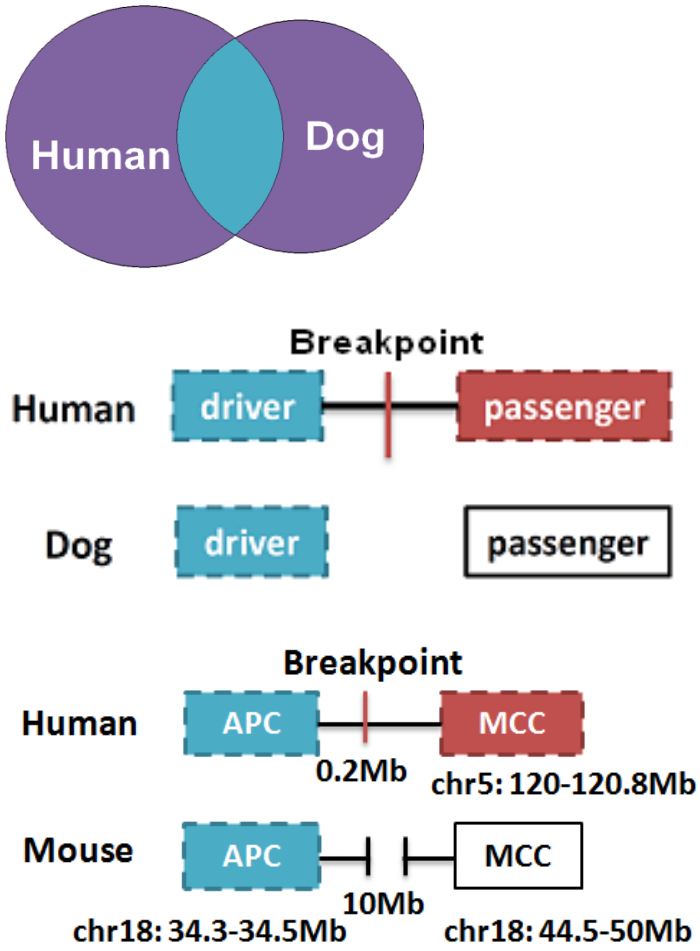

Human
Dog
Breakpoint
Human
driver
passenger
Dog
driver
passenger
Breakpoint
Human
APC
MCC
0.2Mb
chr5: 120-120.8Mb
Mouse
APC
MCC
10Mb
chr18: 34.3-34.5Mb
chr18: 44.5-50Mb

Figure 1.2

REFERENCES

Altekruse SF, McGlynn KA, Reichman ME. 2009. Hepatocellular carcinoma incidence, mortality, and survival trends in the United States from 1975 to 2005. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 27(9): 1485-1491.

Boveri T. 2008. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Journal of cell science 121 Suppl 1: 1-84.

Bray GA, Smith SR, de Jonge L, Xie H, Rood J, Martin CK, Most M, Brock C, Mancuso S, Redman LM. 2012. Effect of dietary protein content on weight gain, energy expenditure, and body composition during overeating: a randomized controlled trial. JAMA : the journal of the American Medical Association 307(1): 47-55.

Bruce WR, Van Der Gaag H. 1963. A Quantitative Assay for the Number of Murine Lymphoma Cells Capable of Proliferation in Vivo. Nature 199: 79-80.

Buchkovich K, Duffy LA, Harlow E. 1989. The retinoblastoma protein is phosphorylated during specific phases of the cell cycle. Cell 58(6): 1097-1105.

Cappuzzo F, Hirsch FR, Rossi E, Bartolini S, Ceresoli GL, Bemis L, Haney J, Witta S, Danenberg K, Domenichini I et al. 2005. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. Journal of the National Cancer Institute 97(9): 643-655.

Cavenee WK, Dryja TP, Phillips RA, Benedict WF, Godbout R, Gallie BL, Murphree AL, Strong LC, White RL. 1983. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 305(5937): 779-784.

Cheon DJ, Orsulic S. 2011. Mouse models of cancer. Annual review of pathology 6: 95-119.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 6(1): 99-103.

Choate KA, Lu Y, Zhou J, Choi M, Elias PM, Farhi A, Nelson-Williams C, Crumrine D, Williams ML, Nopper AJ et al. 2010. Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in KRT10. Science 330(6000): 94-97.

Comings DE. 1973. A general theory of carcinogenesis. Proceedings of the National Academy of Sciences of the United States of America 70(12): 3324-3328.

DeCaprio JA, Ludlow JW, Lynch D, Furukawa Y, Griffin J, Piwnica-Worms H, Huang CM, Livingston DM. 1989. The product of the retinoblastoma susceptibility gene has properties of a cell cycle regulatory element. Cell 58(6): 1085-1095.

DeVita VT, Jr., Chu E. 2008. A history of cancer chemotherapy. Cancer research 68(21): 8643-8653.

Diamandopoulos GT. 1996. Cancer: an historical perspective. Anticancer research 16(4A): 1595-1602.

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455(7216): 1069-1075.

Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301(5895): 89-92.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. 2010. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer Journal international du cancer 127(12): 2893-2917.

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). Current protocols in human genetics / editorial board, Jonathan L Haines [et al] Chapter 10: Unit 10 11.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nature reviews Cancer 4(3): 177-183.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537.

Gondo Y, Fukumura R, Murata T, Makino S. 2009. Next-generation gene targeting in the mouse for functional genomics. BMB reports 42(6): 315-323.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C et al. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446(7132): 153-158.

Greenwald P, Dunn BK. 2009. Landmarks in the history of cancer epidemiology. Cancer research 69(6): 2151-2162.

Greger V, Passarge E, Hopping W, Messmer E, Horsthemke B. 1989. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. Human genetics 83(2): 155-158.

Guastadisegni C, Colafranceschi M, Ottini L, Dogliotti E. 2010. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. Eur J Cancer 46(15): 2788-2798.

Hajdu SI. 2011. A note from history: landmarks in history of cancer, part 1. Cancer 117(5): 1097-1102.

Howlader N, Ries LA, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. 2010. Improved estimates of cancer-specific survival rates from population-based data. Journal of the National Cancer Institute 102(20): 1584-1598.

Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. 2007. Cancer statistics, 2007. CA: a cancer journal for clinicians 57(1): 43-66.

Ji X, Tang J, Halberg R, Busam D, Ferriera S, Pena MM, Venkataramu C, Yeatman TJ, Zhao S. 2010. Distinguishing between cancer driver and passenger gene alteration candidates via cross-species comparison: a pilot study. *BMC cancer* **10**: 426.

Ji X, Zhao S. 2008. DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. Genomics 91(3): 249-258.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A et al. 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321(5897): 1801-1806.

Kallioniemi A. 2008. CGH microarrays and cancer. Current opinion in biotechnology 19(1): 36-40.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258(5083): 818-821.

Karlsson EK, Lindblad-Toh K. 2008. Leader of the pack: gene mapping in dogs and other model organisms. Nature reviews Genetics 9(9): 713-725.

Kastan MB, Onyekwere O, Sidransky D, Vogelstein B, Craig RW. 1991. Participation of p53 protein in the cellular response to DNA damage. Cancer research 51(23 Pt 1): 6304-6311.

Khanna C, Lindblad-Toh K, Vail D, London C, Bergman P, Barber L, Breen M, Kitchell B, McNeil E, Modiano JF et al. 2006. The dog as a cancer model. Nature biotechnology 24(9): 1065-1066.

Kim M, Gans JD, Nogueira C, Wang A, Paik JH, Feng B, Brennan C, Hahn WC, Cordon-Cardo C, Wagner SN et al. 2006. Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. Cell 125(7): 1269-1281.

Kinzler KW, Vogelstein B. 1996. Lessons from hereditary colorectal cancer. Cell 87(2): 159-170.

Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. American journal of human genetics 76(1): 8-32.

Klopocki E, Mundlos S. 2011. Copy-number variations, noncoding sequences, and human phenotypes. Annual review of genomics and human genetics 12: 53-72.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069): 803-819.

Macconaill LE, Garraway LA. 2010. Clinical implications of the cancer genome. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 28(35): 5219-5228.

Martin GS. 1970. Rous sarcoma virus: a function required for the maintenance of the transformed state. Nature 227(5262): 1021-1023.

Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, Antonarakis SE, Reymond A. 2006. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. American journal of human genetics 79(2): 332-341.

Paoloni M, Khanna C. 2008. Translation of new cancer treatments from pet dogs to humans. Nature reviews Cancer 8(2): 147-156.

Peeper D, Berns A. 2006. Cross-species oncogenomics in cancer gene identification. Cell 125(7): 1230-1233.

Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America 100(13): 7672-7677.

Pontius JU, Mullikin JC, Smith DR, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B, Volfovsky N et al. 2007. Initial sequence and comparative analysis of the cat genome. Genome research 17(11): 1675-1689.

Rowell JL, McCarthy DO, Alvarez CE. 2011. Dog models of naturally occurring cancer. Trends in molecular medicine 17(7): 380-388.

Satzinger H. 2008. Theodor and Marcella Boveri: chromosomes and cytoplasm in heredity and development. Nature reviews Genetics 9(3): 231-238.

Sharpless NE, Depinho RA. 2006. The mighty mouse: genetically engineered mouse models in cancer drug development. Nature reviews Drug discovery 5(9): 741-754.

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N et al. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314(5797): 268-274.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes, chromosomes & cancer 20(4): 399-407.

Spector DH, Varmus HE, Bishop JM. 1978. Nucleotide sequences related to the transforming gene of avian sarcoma virus are present in DNA of uninfected vertebrates. Proceedings of the National Academy of Sciences of the United States of America 75(9): 4102-4106.

Stankiewicz P, Beaudet AL. 2007. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. Current opinion in genetics & development 17(3): 182-192.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annual review of medicine 61: 437-455.

Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C et al. 2005. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature genetics 37(6): 590-592.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813): 848-853.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458(7239): 719-724.

Tang J, Le S, Sun L, Yan X, Zhang M, Macleod J, Leroy B, Northrup N, Ellis A, Yeatman TJ et al. 2010. Copy number abnormalities in sporadic canine colorectal cancers. Genome research 20(3): 341-350.

Torkamani A, Schork NJ. 2009. Identification of rare cancer driver mutations by network reconstruction. Genome research 19(9): 1570-1578.

Urban AE, Korbel JO, Selzer R, Richmond T, Hacker A, Popescu GV, Cubells JF, Green R, Emanuel BS, Gerstein MB et al. 2006. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America 103(12): 4534-4539.

Wagner RP. 1999. Anecdotal, historical and critical commentaries on genetics. Rudolph Virchow and the genetic basis of somatic ecology. Genetics 151(3): 917-920.

Wong KM, Hudson TJ, McPherson JD. 2011. Unraveling the genetics of cancer: genome sequencing and beyond. Annual review of genomics and human genetics 12: 407-430.

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J et al. 2007. The genomic landscapes of human breast and colorectal cancers. Science 318(5853): 1108-1113.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. 2006. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic acids research 34(2): 445-450.

Zender L, Spector MS, Xue W, Flemming P, Cordon-Cardo C, Silke J, Fan ST, Luk JM, Wigler M, Hannon GJ et al. 2006. Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. Cell 125(7): 1253-1267.

Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. Genome research 14(10A): 1851-1860.

CHAPTER 2

COPY NUMBER ABNORMALITIES IN SPORADIC CANINE COLORECTAL CANCERS[1]

[1]Tang J, Le S, Sun L, Yan X, Zhang M, MacLeod J, LeRoy B, Northrup N, Ellis A, Yeatman TJ, Liang Y, Zwick ME, Zhao S. 2010. *Genome Research*. 20(3):341-50

ABSTRACT

Human colorectal cancer (CRC) is one of the better-understood systems for studying the genetics of cancer initiation and progression. To develop a cross-species comparison strategy for identifying CRC causative gene or genomic alterations, we performed array comparative genomic hybridization (aCGH) to investigate copy number abnormalities (CNAs), one of the most prominent lesion types reported for human CRCs, in 10 spontaneously occurring canine CRCs. The results revealed for the first time a strong degree of genetic homology between sporadic canine and human CRCs. First, we saw that between 5 and 22% of the canine genome was amplified/deleted in these tumors, and that, reminiscent of human CRCs, the total altered sequences directly correlated to the tumor's progression stage, origin, and likely microsatellite instability status. Second, when mapping the identified CNAs onto syntenic regions of the human genome, we noted that the canine orthologs of genes participating in known human CRC pathways were recurrently disrupted, indicating that these pathways might be altered in the canine CRCs as well. Lastly, we observed a significant overlapping of CNAs between human and canine tumors, and tumors from the two species were clustered according to the tumor subtypes but not the species. Significantly, compared with the shared CNAs, we found that species-specific (especially human-specific) CNAs localize to evolutionarily unstable regions that harbor more segmental duplications and interspecies genomic rearrangement breakpoints. These findings indicate that CNAs recurrent between human and dog CRCs may have a higher probability of being cancer-causative, compared with CNAs found in one species only.

The aCGH data have been submitted to the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE19318.

Running Title: Copy Number Abnormalities in Dog Colon Tumors

Key Words: Canine Colorectal Cancer; Copy Number Aberrations; Cross-Species Comparison; Cancer Driver Identification

INTRODUCTION

Cancer is a disease of the genome, and genomic instability is a hallmark of cancer (Hanahan and Weinberg 2000). As cancer progresses, more extensive genomic instability develops, and more abnormal changes accumulate on the genome (e.g., copy number abnormalities or CNAs, translocations, and inversions) (Albertson et al. 2003). While some of these aberrations disrupt normal cellular processes and indeed contribute to cancer development and progression, others emerge simply as passenger alterations of cancer genomic instability and play no role in disease etiology. Clearly, finding genomic abnormalities is important, but identifying those that are cancer-causative is even more meaningful.

A central aim of cancer research has been to identify causative (or driver) alterations. This has become both increasingly challenging and urgent in recent years with the launch of high-throughput cancer genome projects, such as the Cancer Genome Atlas (cancergenome.nih.gov/components/cgcc.asp), the International Cancer Genome Consortium (www.icgc.org/), and others (e.g., Greenman et al. 2007; Jones et al. 2008). Researchers have been tackling this challenge by improving experimental conditions (e.g., high-resolution microarray to refine the boundaries of amplicons to narrow down the "driver" genes; see Haverty et al. 2008) and by developing more sophisticated statistical models and functional analysis strategies to systematically (Aebersold et al. 2009) identify significant abnormalities (e.g., Greenman et al. 2007; Jones et al. 2008; Beroukhim et al. 2007).

We are developing a cross-species comparison strategy that differs fundamentally from the current published approaches described above (which study humans only). We hypothesize that causative alteration candidates can be distinguished from consequent candidates by examining orthologous genes or genomic loci with tumors from multiple species having the same type of cancer. Provided these species share similar molecular and genetic pathways of cancer development and progression, abnormalities that are recurrent among different species will be deemed causative, whereas those that are found in only one species and are located in evolutionarily unstable sites will be considered consequent (bystanders or passengers) (Figure 2.1). In our study, evolutionarily unstable sites are defined as regions enriched with interspecies genomic rearrangement breakpoints (ISGRBPs) (Zhao et al. 2004; Pevzner & Tesler 2003) and segmental duplications (SDs) (Bailey et al. 2002 & 2004; Kidd et al. 2008; Turner et al. 2007).

Our hypothesis rests on the same rationale that cancer researchers have been using for years: abnormalities recurrent among different cases are more likely to be causative, compared with non-recurrent events. The difference is we are searching for events that are recurrent not only among different cases within the same species, but also among different species. Compared with single-species approaches, this multi-species comparison strategy can better distinguish causative events from passenger alterations of cancer genomic instability, by taking account of the difference in the genomic location of orthologous genes and loci caused by interspecies genomic rearrangements that occurred during evolution (Figure 2.1). For instance, chromosome region 18q is often found to be deleted in human colorectal cancer (CRC). Except for *SMAD4*, which is clearly

demonstrated to be a CRC driver gene (Kinzler and Vogelstein 1996), the roles of many 18q genes in CRC development and progression remain unclear, and it is possible that some of these genes are deleted in human CRC simply because they are near *SMAD4*. We have found that interspecies genomic rearrangements have dispatched many such genes far away from *SMAD4* via translocations or inversions in the dog genome. Studying dog CRCs could shed light on whether these genes are drivers or passengers, provided we can first show that dog and human CRCs share similar molecular and genetic pathways of cancer development and progression.

To test this cross-species strategy, we have been conducting a comparative study between sporadic human and canine CRCs. Human CRC is one of the better-understood systems for studying the genetics of cancer initiation and progression. The proposed stepwise model of human colorectal carcinogenesis (Kinzler & Vogelstein 1996; Rajagopalan et al. 2003) highlights the key role of genomic instability, which occurs in the form of either chromosomal instability (CIN) (Lengauer et al. 1997) or microsatellite instability (MSI) (Toft and Arends 1998). MSI, which is characterized by a high level of single- or oligo-base mutations due to defective DNA mismatch repair (MMR) (Grady 2004) and recognized by indel mutations in microsatellite loci (Boland et al. 1998), occurs in ~13% of human sporadic CRCs. CIN, characterized by CNAs of relatively narrow genomic regions ranging from single loci to entire chromosomes, occurs in 87% of human sporadic CRCs. Although the role that CIN plays in causing cancer is unclear, the occurrence of CNAs could result in inactivation of tumor suppressors (e.g., *APC*, *SMAD4* and *TP53*), as well as overactivation of oncogenes, such as *KRAS*. This could in

turn disrupt key signaling pathways of Wnt, TGF-β, p53 and others that play fundamental roles in CRC development and progression (Kaiser et al. 2007; Sancho et al. 2004; Grady 2004; Rajagopalan et al. 2003; Gallahan & Callahan 1997; Kinzler & Vogelstein 1996).

Because of its importance in human CRCs, we investigated genomic instability in 10 spontaneously occurring dog CRCs and reported the results here. This study yielded for the first time several pieces of evidence in support of dog CRCs possibly following similar molecular pathways of cancer development and progression as human CRCs. First, via array comparative genome hybridization (aCGH) analyses, we found that reminiscent of human CRCs, between 5 to 22% of the canine genome was either amplified or deleted in these canine tumors, and that the total amount of altered genomic sequences directly correlated to the tumor's progression stage, origin, and likely MSI status. Second, mapping of the identified dog CNAs to the better-annotated human genome revealed that the canine orthologs of many genes involved in known human CRC development and progression pathways were disrupted, indicating that these pathways might also be altered in these canine CRCs. Third, we observed a significant overlapping of CNAs between human and canine tumors, and tumors from the two species were clustered according to the tumor subtypes, but not according to the species. Besides these genetic similarities between the two species, our analyses revealed that the CNAs found in only one species (especially the human) localize to evolutionarily unstable genomic regions that harbor more SDs and ISGRBPs compared with CNAs found in both species. This indicates that shared CNAs might be more significant in CRC etiology compared with species-specific (especially human-specific) CNAs.

RESULTS

CNA identification of dog colorectal tumors

We performed aCGH analyses on a total of 10 sporadic dog colorectal tumors, which covered tumors at early (adenomas) and late (adenocarcinomas) cancer progression stages (Figure 2.2), as well as tumors of epithelial origin (adenomas and adenocarcinomas) and non-epithelial origin (a mast cell tumor; a leiomyosarcoma which arose from muscle cells). As summarized in Table 2.1, we identified between 2,000 and 10,000 CNAs per canine tumor genome, with sizes ranging from 17 kb to 1.7 Mb with an average of 40-60 kb (varied with the tumors; see Figure 2.3 for the identified CNAs for one tumor). These CNAs have a total probe number of between 5 and 208, with $\log_2$-ratio means ranging from 0.15 to 2.36. To validate our aCGH analyses, we performed quantitative polymerase chain reaction (qPCR) analyses on a small fraction of CNAs identified, including regions within a number of genes (e.g., *GSK3B*, *APC*, *SMAD3*, *SMAD2*, *SMAD4*, *TGFB2*, *TGFBR2*, *PTEN*, and *TP53*) as well as two deleted regions on chromosomes 5 and 9. Through t-tests (see Materials and Methods), we confirmed deletions in *APC*, *PTEN*, *SMAD3*, *SMAD2*, *SMAD4*, *TP53*, and the regions on chromosomes 5 and 9, as well as amplifications in *TGFB2* and *GSK3B* (p-values ranged from 0.001 to <0.1), indicating the accuracy of our aCGH analyses.

As shown in Table 2.1, the total amount of CNAs identified varied significantly among the ten tumor genomes; as a result, the amplified/deleted regions ranged between 5 and 22% of the genome. Reminiscent of human CRCs, we found that this variation correlated well with the progression stage, the origin, and the likely the MSI status of the

tumors. First, tumors with fewer CNAs are at early stages (~146-285Mb for adenomas), whereas tumors with a greater number of CNAs are at late stages (~350-550Mb for adenocarcinomas) (Table 2.1). Second, our study revealed that, for tumors of similar stages, those with an epithelial origin harbor more CNAs than those arising from smooth muscle cells (138 Mb) and mast cells (197 Mb). Finally, our analyses indicate that the presence of MSI is associated with fewer CNAs. As shown in Figure 2.4, one of the eight microsatellite loci was found altered in an adenoma, which therefore should be classified as MSI-low. Even though this adenoma is more advanced than the other two adenomas (which are MSI-none) based on the histopathological properties, it has the fewest CNAs (146Mb vs. 285 and 253Mb).

The correlation described above was more quantitatively measured with a multiple linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, where ε_i is the random error, and y_i, x_{i1}, x_{i2} and x_{i3} represent the total CNA amount, stage, cellular origin, and MSI status of the i_{th} tumor ($i = 1,2,3...,10$), respectively. We found that the tumor T3 (Table 2.1) is an outlier in this analysis; if excluding T3, the observed data were predicted by $\hat{y} = -226.5 + 271x_1 + 372.5x_2 + 123x_3$, with a coefficient of determination $R^2 = 0.99$ (the adjusted $R^2 = 0.98$) and $p < 0.0001$. The analysis also confirmed that the amount of CNAs was indeed significantly related to the tumor stage ($p < 0.0001$), the cellular origin ($p < 0.0001$), and the MSI status ($p = 0.01$). We understand that we had a small sample size here (only 9 tumors). According to Knofczynski (2008), the sample size requirement of multiple linear regression varies with R^2 and, for $R^2 > 0.9$, a good

predication with 3-variables requires at least 9 samples. Hence, our analysis satisfied this minimal sample size requirement.

We also performed clustering analyses using pair-wise overlapping CNAs. Both the minimum spanning tree (MST) (Cormen et al. 2001; Kruskal 1956; Prim 1957) and top-down clustering strategies (Liu et al. 2006) (see Materials and Methods), as well as hierarchical clustering (Hastie et al. 2009), yielded the same structures as shown in Figure 2.5. Reminiscent of human CRCs, the dog tumors of similar subtypes were clustered together. For instance, the late-stage tumors were grouped together, splitting from early-stage tumors in the tree. In addition, tumors with non-epithelial origins or those that are likely to be MSI-low are separate from those that arose from colon epithelial cells and were MSI-none. We found that grouping of subtypes in the cluster was mostly caused by the differences in the total CNA size among the subtypes, consistent with the multiple linear regression results described above. However, within the same subtype of adenocarcinomas, tumors clustered based on not only the total CNA size but also the CNA pattern.

Genes and pathways disrupted in the dog colorectal tumors

The human genome is better annotated than the dog genome. Consequently, to more accurately determine which genes are affected by the CNAs identified above, we mapped the dog CNAs onto the orthologous sites of the human genome. This was done by converting the dog genomic sequence coordinate of each CNA to the corresponding human coordinate, using a high-resolution human-dog genomic synteny map anchored by

bacterial artificial chromosome (BAC) clones that we previously built (Zhao et al. 2004; Ji and Zhao 2008). Depending on the tumors, between 97.3% and 98.2% of the CNAs were successfully placed onto the human genome (Table 2.1; see Figure 2.3 for one example). The remaining unmapped CNAs were either found to fall in the rearrangement breakpoint regions (74-88%) or were within or near the telomeric regions where the human-dog synteny is not yet resolved (12-26%). This precluded their accurate localization on the human genome. Importantly, the unmapped CNAs do not encode genes, and thus they will not affect the analysis results described below.

Using the mapping information described above, we found about 2,000-4,000 genes for early-stage/non-epithelial/probable MSI-low tumors and 5,000-7,500 genes for the late stage epithelial tumors (or adenocarcinomas) (Table 2.1). For 73-86% of these genes, we were able to annotate their functions and classify them using the GO (gene ontology) system at the biological process level. This analysis assigned signal transduction, transcription, cell differentiation, cell death or proliferation, and/or other biological functions to these genes.

Importantly, we asked if any of these genes participate in pathways known to be involved in human CRC development and progression (Kaiser et al. 2007; Sancho et al. 2004; Grady 2004; Rajagopalan et al. 2003; Gallahan & Callahan 1997; Kinzler & Vogelstein 1996). These included well-characterized signaling pathways of Wnt, TGF-β, p53, and MAPK, as well as pathways controlling cell growth and apoptosis, and cell cycle (see www.genome.jp/kegg/pathway/hsa/hsa05210.html). Among the 647 total

genes included in these pathways, we found that significantly more genes were disrupted in the dog tumors than predicted by the random model (p=0.0004), which assumed that overlaps between disrupted genes and the pathway members were totally random (Table 2.2a). Table 2.2b lists those genes that were recurrently disrupted among the dog tumors, identified following a strategy described by Beroukhim et al. (2007) (see Materials and Methods). These results indicate that the human CRC pathways were also likely to be disrupted in dog CRCs, demonstrating the genetic and molecular similarities of CRCs between these two species.

Comparing dog CNAs to human CNAs

Using Roche NimbleGen's 2.1-million human oligo arrays, we performed aCGH analyses on a human adenocarcinoma (a Dukes B tumor) that is MSI-none, as well as another adenocarcinoma (also a Dukes B tumor) that is MSI-high (at least two MSI loci disrupted). Consistent with the literature reports, we found significantly more CNAs in the MSI-none tumor than the MSI-high tumor (312 Mb vs. 20.5 Mb). In addition, for the MSI-none tumor, we identified amplifications in the 5p, 8q, 13q, and 20q regions, as well as deletions in the 8p, 11q, 14q, 15q, 17p, 18q, and 21q regions. Those are consistent with previous aCGH studies (e.g., Camps et al. 2006; Douglas et al. 2004; and Nakao et al. 2004). Like the dog analyses described before, these results also demonstrate the accuracy of our aCGH analyses.

Importantly, we found that approximately 12% to 33% of CNAs of the human MSI-none tumor co-localized with the dog CNAs on the human genome, which are

significantly higher than those predicted (3-8%) by assuming that the human tumors and the dog tumors are completely unrelated and their CNA overlap is totally random ($p<10^{-4}$). This close relationship was also demonstrated by the clustering analyses. As shown in Figure 2.5b, the human MSI-none adenocarcinoma was near the canine MSI-none adenocarcinomas, whereas the human MSI-high tumor was near the canine MSI-low tumor (although this grouping was mostly caused by the difference in the total CNA size, similar to the dog tumors described previously). Thus, the tumors were clustered according to the tumor subtype, but not according to the species (Figure 2.5b). These results provide another piece of evidence demonstrating the molecular and genetic similarity between human and canine CRCs.

To better differentiate the CNAs shared between the two species from species-specific CNAs, we investigated the evolutionary genomic instability of these regions by determining the amount of interspecies genomic rearrangement breakpoints (ISGRBPs) (mouse, rat, and dog genomic rearrangement breakpoints on the human genome) and segmental duplications (SDs) in them. We found that the shared CNAs contain 2-3 times fewer ISGRBPs and up to 3 times fewer SDs compared with the human-specific CNVs. To ensure that this enrichment was not caused by mapping failures, we mapped the human CNAs onto the dog genome using the same synteny file described previously. We found that human-specific CNAs were mapped with same efficiency as the total CNAs, indicating that the enrichment was unlikely an artifact caused by mapping issues.

To further expand this analysis by including more human tumors, we searched literature and databases for findings reported by other groups. We found that such studies were mostly performed with BAC arrays (e.g., Camps et al. 2008) but not with high-density oligo arrays such as we applied here. We thus focused on the BAC data and extracted the CNA information from 53 human CRC tissue samples reported in the NCBI's SKY/M-FISH and CGH database (Knutsen et al. 2005), selecting those CNAs that were recurrent among ≥11 tumors (≥20% of the total sample size). Unfortunately, the human CNAs were reported in chromosomal cytogenetic bands only. Consequently, we converted the corresponding human sequence coordinates of the dog CNAs that had been successfully mapped onto the human genome (see above) into the smallest possible cytogenetic bands. Then, we identified those cytogenetic bands where ≥20% of their genomic sequences were amplified or deleted in at least two dog tumors (≥20% of the total sample size) to compare with the selected human CNAs.

As summarized in Table 2.3, we found that approximately 64% of the human CNAs and 56% of the dog CNAs co-localize in the genome. These included well-known aberration sites in human CRCs such as those in 7p, 7q, 17p, 17q, and 18q. In addition, we found that the ISGRBP density was 3.66×10^{-7} per base for the shared CNAs, 5.16×10^{-7} per base for the human-specific CNVs, and 4.58×10^{-7} per base for the dog-specific CNVs. Similarly, the shared CNVs were found to harbor fewer SDs (2%) compared with those that are human-specific (5.4%) or dog-specific (2.9%). Thus, consistent with the results of the oligo arrays described above, species-specific CNAs locate in evolutionarily more unstable regions compared with shared CNAs.

DISCUSSION

Animal cancer models (rodents, zebrafishes, dogs, cats, and others) have been used extensively in human cancer research (e.g., Maser et al. 2007; Hinoi et al. 2007; Hansen & Khanna 2004; Rosol et al. 2003; Vail & MacEwen 2000; Moser et al. 1993). Sporadic canine cancers should make excellent models for studying the corresponding human cancers for a number of reasons. First, they are naturally occurring and heterogeneous, unlike most genetically modified or xenograft rodent cancer models. Second, there are numerous anatomic and clinical similarities between humans and dogs with the same type of cancer (Paoloni & Khanna 2008 & 2007; Khanna et al. 2006; Vail & MacEwen 2000; Michell 2004; Argyle 2009 & 2005; LeRoy & Northrup 2009). Third, the dog genome has been sequenced, and an accurate version of the dog genomic sequence assemblies is available (Lindblad-Toh et al. 2005), facilitating many experimental and bioinformatics analyses. For these reasons, we have focused on using sporadic dog CRCs to develop a novel cross-species comparative genomics and oncology strategy to identify causative CRC alteration candidates (Figure 2.1).

Consistent with the reported anatomic and clinical similarities, our initial characterization of genomic instability in ten spontaneously occurring dog tumors revealed for the first time the genetic and molecular similarities between sporadic human and dog CRCs. All ten canine tumors investigated exhibited CIN, with the extent of CIN correlating with the tumor's progression stage, origin, and likely MSI status reminiscent of human CRCs. However, the MSI phenotype was found in only one canine tumor. This is consistent with human sporadic CRCs, where the majority (87%) display CIN and the

minority (13%) display MSI. In addition, our analyses indicate the corresponding human CRC pathways may also be altered in the dog cancers and that tumors from the two species were clustered according to the tumor subtypes, but not according to the species. These observations indicate that human and dog CRCs might share similar molecular and genetic pathways of cancer development and progression. This possibility is further supported by ongoing analyses that focus on investigating the expression alteration of bona fide CRC genes such as *APC*, *SMAD4*, and *TP53* in the dog tumors (their expression indeed altered by quantitative reverse transcription PCR analyses, p-values ranged from 0.001 to <0.1). Of course, more studies are needed. For instance, for the dog MSI study, the eight dog microsatellite loci (Figure 2.4), of which five are homologs of standard human MSI markers (Boland et al. 1998) and three are dinucleotide markers used to characterize MSI in canine mammary gland neoplasia (McNiel et al. 2007), have not been extensively tested. Importantly, we do not know whether disruption of the dog locus shown in Figure 2.4, which belongs to one of three dinucleotide markers, is due to defective MMR or not. Hence, it would be revealing to examine the MMR system in the probable MSI+ dog tumor. It would also be informative to conduct expression microarray analyses to identify additional genes and pathways that are altered in dog CRCs, similar to those performed with mouse tumors by Kaiser et al. (2007). Of course, more dog tumors merit investigation.

We discovered that the species-specific CNAs localize to genomic regions that are evolutionarily more unstable (having more ISGRBPs and SDs) compared with the CNAs shared between the two species. For instance, although the human genomic region

8p23.1 (6.2-12.7Mb) was found to be frequently disrupted in human CRCs (e.g., Camps et al. 2008), we did not find any changes in the corresponding region in the ten dog tumors investigated (Table 2.3). Various studies have reported that 8p23.1 belongs to an evolutionarily hypervariable region, enriched with SDs and inversions that are specific to apes. For example, we recently identified 9 complete/truncated copies of a 300 kb LTR-retrotransposon-like element clustered in this region; these copies have likely facilitated inversions observed among different primate species (Ji and Zhao 2008) and even among different normal human individuals (Deng et al. 2008). Interestingly, a recent sequencing effort reported that, in the genome of the breast cancer cell line MCF7, dispersed rearrangement breakpoint regions are significantly enriched with SDs (5.2X) whereas clustered breakpoints regions are not (Hampton et al. 2009). It would be interesting to know if clustered breakpoint regions harbor more cancer-driver alterations than dispersed rearrangement breakpoint regions.

Our finding raises the possibility that the human-specific CNAs may be a consequence of cancer development and progression, rather than a cause. This is because these genomic regions are intrinsically more unstable and are thus more prone to changes when the genome becomes increasingly more unstable as cancer progresses, compared with other genomic sites. Thus, these human-specific CNAs may be less significant than the shared CNAs in CRC etiology, especially if both species are clearly shown to follow similar molecular and genetic pathways of cancer development and progression. Certainly more dog CRCs should be investigated if we are to more accurately define the shared- and species-specific CNA sets. Although much more work still needs to be done,

this study demonstrated the promise of using sporadic dog CRCs to identify bona fide CRC gene and other genomic abnormality candidates. However, to effectively achieve this goal, we must emphasize that significantly more dog and human tumors should be compared and the identified driver alteration candidates should be verified with further experimental studies such as those described by Yang et al. (2008).

MATERIALS AND METHODS

Canine colorectal tissue samples.

Frozen samples of canine colorectal tumors and normal colon tissues were obtained from the Veterinary Teaching Hospital of the University of Georgia College of Veterinary Medicine, the William R. Pritchard Veterinary Medical Teaching Hospital of the University of California-Davis School of Veterinary Medicine, as well as the Animal Cancer Tissue Repository at the Colorado State Universality. Tissue samples were obtained during surgery. After washing in phosphate-buffered saline, they were snap-frozen in liquid nitrogen for 10 minutes and then stored at -80 ℃ until further analyses.

Tissue sample cryomicrodissection and DNA extraction.

With a cryostat, we first cut a tissue sample into two or three pieces, depending upon its size; then from the fresh-cut side of the largest piece, we sectioned typically three 10-micron slices, which were stained with hematoxylin and eosin Y (Figure 2.2). Based on the staining results, we used a surgical knife to dissect the sections from the tissues that are enriched with tumor cells to minimize normal cell contamination or normal colon epithelial cells to minimize other types of cells such as cells of muscle, fat,

and connective tissues. Finally, genomic DNA was extracted from the dissected tissues using a Qiagen Tissue DNeasy purification kit.

MSI Assay.

MSI analysis was performed with 5 dog microsatellite markers that share high sequence homology with corresponding human microsatellite markers used to test MSI in human CRCs (Boland et al. 1998): mononucleotides dBAT25 and dBAT26, and dinucleotides dD8S87, dD17S787, and dD20S100. Further, we included 3 additional dinucleotide markers (CPH14, CPH5, and AHTK209) based on a previous study of canine mammary gland neoplasia (McNiel et al. 2007).

PCR reactions were carried out in a 10-µl reaction volume containing 25 ng of genomic DNA, 25 µM of primers, and 5 µl of 2X iQ^{TM} supermix (Bio-Rad). The reactions were performed under the following conditions: 95 ℃ for 5 min, 30 cycles of 94 ℃ for 30 sec, 55 ℃ for 30 sec, and 72 ℃ for 30 sec, and a final extension at 72 ℃ for 10 min. The PCR products were resolved on a 10% polyacrylamide denaturing gel containing 5.7 M urea and visualized by silver staining using a Silver Staining Plus kit from Bio-Rad.

Human DNA samples purified from paired normal/tumor tissues and associated pathology information were provided by Dr. Timothy J. Yeatman from H. Lee Moffitt Cancer Center and Research Institute in Florida. The MSI assay was performed with

standard microsatellite loci described by Boland et al. (1998) following the procedure described above.

Dog array comparative genomic hybridization (aCGH).

The aCGH hybridization, as well as data collection and initial analysis, were conducted at Dr. Michael E. Zwick's laboratory at Emory University, following the standard protocol contained within the CGH "NimbleChip^{TM} Arrays User's Guide." Test and reference samples were hybridized to standard 385 K canine CGH arrays manufactured by Roche NimbleGen Systems, Inc. Each array was fabricated from a single chip design containing ~ 385,000 probes of ~ 50 bp long oligos selected from unique sequences in the canFam2 genome. This design provides an average resolution of 1 probe every 5-6 kb across the canine genome.

Human aCGH experiments were performed as described above, except that the Roche NimbleGen's human high density arrays were used. Each of these arrays contained ~2.1 million oligo probes, providing a resolution of 1 probe every 1 kb across the human genome, on average.

CNAs were identified by analyzing the $\log_2$-ratios using a software program called SEG, which we recently developed in order to more effectively decipher high density oligo aCGH data for CNA finding. SEG consists of two major steps: change-point identification at the chromosomal level and CNA identification at the whole genome level. For change-point finding, SEG first requires a user-input of the minimal

probe numbers that a CNA should have (which we set to 5 for oligo aCGH data because we relied on signals from at least 5 continuous probes but not individual probes to reduce false positives). With this input, SEG determines the maximum change-points for a chromosome and assigns these initial changes-points. Then, SEG shifts these temporarily-assigned change-points to their correct position by recursively minimizing variations within each segment and removing insignificant change-points by merging neighboring segments where the $\log_2$-ratio means are not significantly different (we set the significance level to 0.01). Once change-points are identified for all chromosomes, SEG uses a false discovery rate (FDR) controlled procedure (Benjamini & Hochberg 1995) to determine which segments are amplified or deleted at the whole genome level. For this study, we set the desired FDR to 0.05, the cutoff total probe number to 5 and the cutoff $\log_2$-ratio mean to 0.25 for CNA identification. The SEG program can be obtained from www.bmb.uga.edu/szhao.

Multiple linear regression analysis was performed using the statistic software JMP8® (www.jmp.com) (SAS Institute, Cary, NC). To perform the analysis, we assigned a value of either 0 or 1 to the tumor stage (0 for adenomas and 1 for adenocarcinomas), the cellular origin (0 for non-epithelial origin and 1 for epithelial origin), and the MSI status (0 for MSI+ and 1 for MSI-) of each tumor.

qPCR analysis.

qPCR reactions were performed in triplicates with each well containing 10 µl iQ™ SYBER Green Supermix from Bio-Rad, 500 nM primers each, and 10ng genomic

DNA in a total reaction volume of 20 μl, with an iCycler iQ Real-Time PCR machine. The amplification condition was: 95 ℃ for 10 seconds, 65 ℃ for 45 seconds, and 78 ℃ for 20 seconds for a total of 40 cycles. The Ct value (the threshold cycles: the number of cycles at which the earliest measurable fluorescence signal can be detected in the qPCR assay; a higher the Ct value means fewer templates) was collected for each reactions. Then, a t-test was conducted to determine if the Ct (normalized based on the total genomic DNA) difference between the tumors and normal samples was significant or not for each gene at a chosen significance level.

Tumor clustering.

Tumors were clustered using the CNAs identified above as follows. For any two tumors T_i and T_j of the ten tumors studied, C_i and C_j represented the total genomic size of CNAs in T_i an T_j respectively, and C_{ij} represented the genomic size of the shared CNAs between T_i and T_j. We defined the similarity between T_i and T_j by $s_{ij} = \frac{C_{ij}}{C_i + C_j - C_{ij}}$ and the distance between T_i and T_j by $d_{ij} = 1 - \frac{C_{ij}}{C_i + C_j - C_{ij}}$.

Minimum spanning tree (MST) tumor clustering. We constructed a weighted complete graph *G* of ten vertices with each vertex representing a tumor and the distance connecting tumors T_i and T_j being d_{ij} calculated above. Then, we applied the Kruskal's algorithm (or the Prim's algorithm) to find the MST of *G* **(**Cormen et al. 2001; Prim 1957; Kruskal 1956). Lastly, we constructed the final tree *T* from the MST by first setting

the root of *T* to represent the cluster harboring all the tumors. Then, starting from this cluster, we recursively divided each cluster into two subclusters using the longest distance between the two tumors within the MST of each cluster until each terminal cluster contained only one tumor.

Top-down clustering. We also clustered the tumors using the top-down clustering algorithm (Steinbach et al. 2000) developed by Liu et al. (2006). First, we performed bisection clustering by initially choosing the two tumors with the lowest similarity as the seeds of two initial clusters and next assigning the remaining tumors to one of these clusters whose seed has a higher similarity to the tumor. Then, we refined these two clusters by recursively moving each tumor from one cluster to another and also exchanging tumors between the clusters. Each movement was evaluated by the two criteria named internal compactness (IC) and internal separation (IS) as defined by $IC = \sum_{r=1}^{2} \frac{\sum_{i<j, T_i, T_j \in C_r} s_{ij}}{b_r}$ and $IS = \frac{\sum_{T_i \in C_1, T_j \in C_2} s_{ij}}{b_1 b_2}$,where C_1, C_2 represents the two clusters which respectively harbor a total number of b_1, b_2 tumors. Since both measures were computed with the pair-wise similarity among the tumors, higher values of *IC* and lower values of *IS* would represent better clustering quality. Consequently, in our current implementation, if $\frac{IC}{IS}$ was larger after the movement, we would keep the new clusters; otherwise, we would discard the movement and keep the original clusters. Similar to how the two initial clusters were established, we recursively applied these bisection clustering and cluster refinement processes to each cluster found in the previous round, until there was only one tumor left in each of the terminal clusters.

Data sources and data integration

The dog aCGH data analyses are based on the canFam version 2.0 and the human genome NCBI build 36.1. Canine CNAs were mapped to the human genome using the BAC clone-based human-dog synteny map that we previously constructed (Zhao et al. 2004; Ji and Zhao 2008). The genes inside the mapped dog CNAs were identified using the KnownGene database downloaded from the University of California Santa Cruz (UCSC) genome site (www.genome.ucsc.edu). Through database cross-linking, functions of the identified genes were annotated based on the GOA database, version 1.109 (www.ebi.ac.uk/GOA/). The subsequent GO slim and classification analysis at the biological process level was achieved by Blast2Go, version 2.35 (www.blast2go.de/). Information on known human CRC-related pathways was obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) release 50.0 (www.genome.jp/kegg). Genes recurrently disrupted in dog tumors (Table 2.2b) were identified by collectively considering the magnitude of a CNA (the value of the $\log_2$-ratios) as well as the percentage of the tumors having this CNA following a procedure described by Beroukhim et al. (2007). Previously published CNAs of human CRCs from BAC aCGH studies were downloaded from the NCBI's SKY/M-FISH and CGH database at www.ncbi.nlm.nih.gov/sky. Human and dog segmental duplication (SD) data, as well as human cytogenetic band and genomic sequence coordinate conversion data, were obtained from the annotation dataset at the UCSC genome sites. Dog, mouse, and rat genomic synteny breakpoint data on the human genome were obtained from previous studies (Zhao et al. 2004; Ji and Zhao 2008).

Acknowledgements.

We are grateful for the initial MSI work performed by Dr. Eric Jiang and Mr. Xueyang Li. We appreciate the help of Dr. Carolyn J Henry of University of Missouri Columbia and Dr. Susan E Lana of Colorado State University in collecting dog samples. The study was funded by the American Cancer Society.

FIGURE LEGENDS

Figure 2.1a Cross-species comparison for causative (or driver) aberration identification. Once we demonstrate that the same types of cancer from the human and the dog share similar molecular and genetic pathways of cancer development and progression, we will consider abnormalities recurrent between the two species as driver candidates (blue area), and those found in only one species and falling in evolutionarily unstable sites (EIN sites) as passenger candidates (red area). Those in the purple area need further studies.

Figure 2.1b The advantage of the human-dog comparison strategy over the human-only strategy for cancer driver gene identification. The cross-species comparison strategy can make use of the difference in the genomic location of orthologous genes between the human and the dog, a result of evolutionary genomic rearrangements that occurred since the two species diverged more than 75 million years ago. The figure shows that two genes, which are nearby in the human genome but distant in the dog genome, are both disrupted in the human cancer (squares with broken lines). However, in the dog cancer, only one gene is disrupted, which will be considered as driver, and the other is intact (square with unbroken lines), which will be deemed as passenger.

Figure 2.2 Cryosectioning and H&E staining of dog colon tumor and normal tissue samples. The images represent a normal colon tissue (top), an adenoma (middle), and an adenocarcinoma (bottom).

Figure 2.3 A: CNAs identified in the genome of a dog adenocarcinoma (T11). The identified CNAs (3,396 gains and 5,711 losses) amount to 551 Mb (22% of the dog genome) (Table 2.1). Each line represents a dog chromosome with its chromosome number indicated on the right. Red/blue vertical lines shown above/below the chromosomes represent gains/losses, respectively, with their length calculated based on $\sqrt{l}\,m$, where *l* and *m* are the total probe number and the mean $\log_2$-ratio of the CNA. Except for CNAs that are larger than 1 Mb in size, the width of the vertical lines is not drawn to scale with the chromosome length. **B: Mapping dog CNAs onto the human genome.** A total of 9,107 CNAs and 541.6 Mb (98.2% of the total) of the same dog tumor shown above were mapped onto the human genome, amounting to 609 Mb on the human genome.

Figure 2.4 MSI assay of the dog tumors. A total of 8 dog microsatellite loci were analyzed as described in the text, with the top five being homologs of frequently used human MSI markers (Boland et al. 1998) and the bottom three being dinucleotide markers used by McNiel et al. (2007) to determine MSI status in canine mammary gland neoplasia. The extra band that the tumor displayed for the locus CPA5 was indicated by an arrow. "T" stands for tumor, and N stands for its matching normal.

Figure 2.5a Clustering of the dog tumors. The tree was constructed by MST or top-down clustering as described in the text (both strategies yielded the same tree), with the sample information for each tumor shown on the right. The number for each branch represents the distance $d(X,Y)$ between the two clusters X and Y involved, calculated

by $d(X,Y) = \frac{\sum_{T_i \in X, T_j \in Y} d_{ij}}{|X||Y|}$,where d_{ij} is the distance between a tumor T_i of cluster X and a tumor T_j of cluster Y calculated as described in the text, and $|X|$ and $|Y|$ are the total number of tumors inside clusters X and Y, respectively.

Figure 2.5b Clustering of tumors from both humans and dogs. The tree was constructed as described above, using the overlapping information of CNAs either identified on (for the human tumors T2551 and T3912) or mapped onto (for the dog tumors, see Figure 2.3) the human genome. MSI-L: MSI-low; MSI-H: MSI-high.

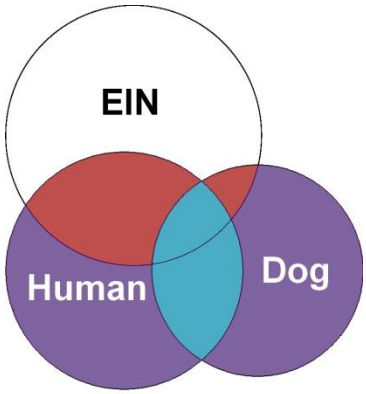


Figure 2.1a

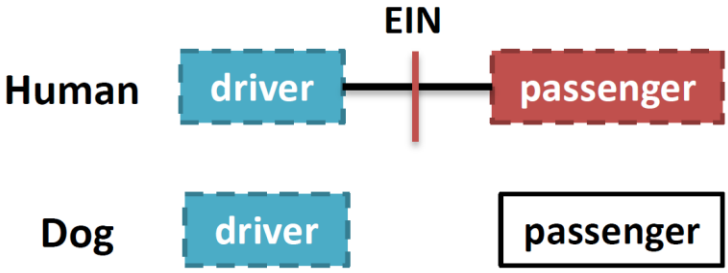


Figure 2.1b

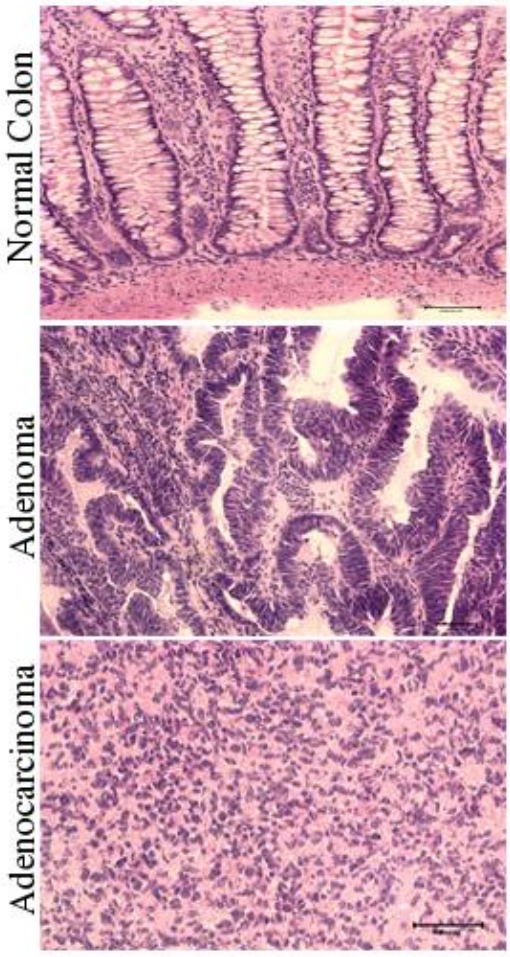


Figure 2.2

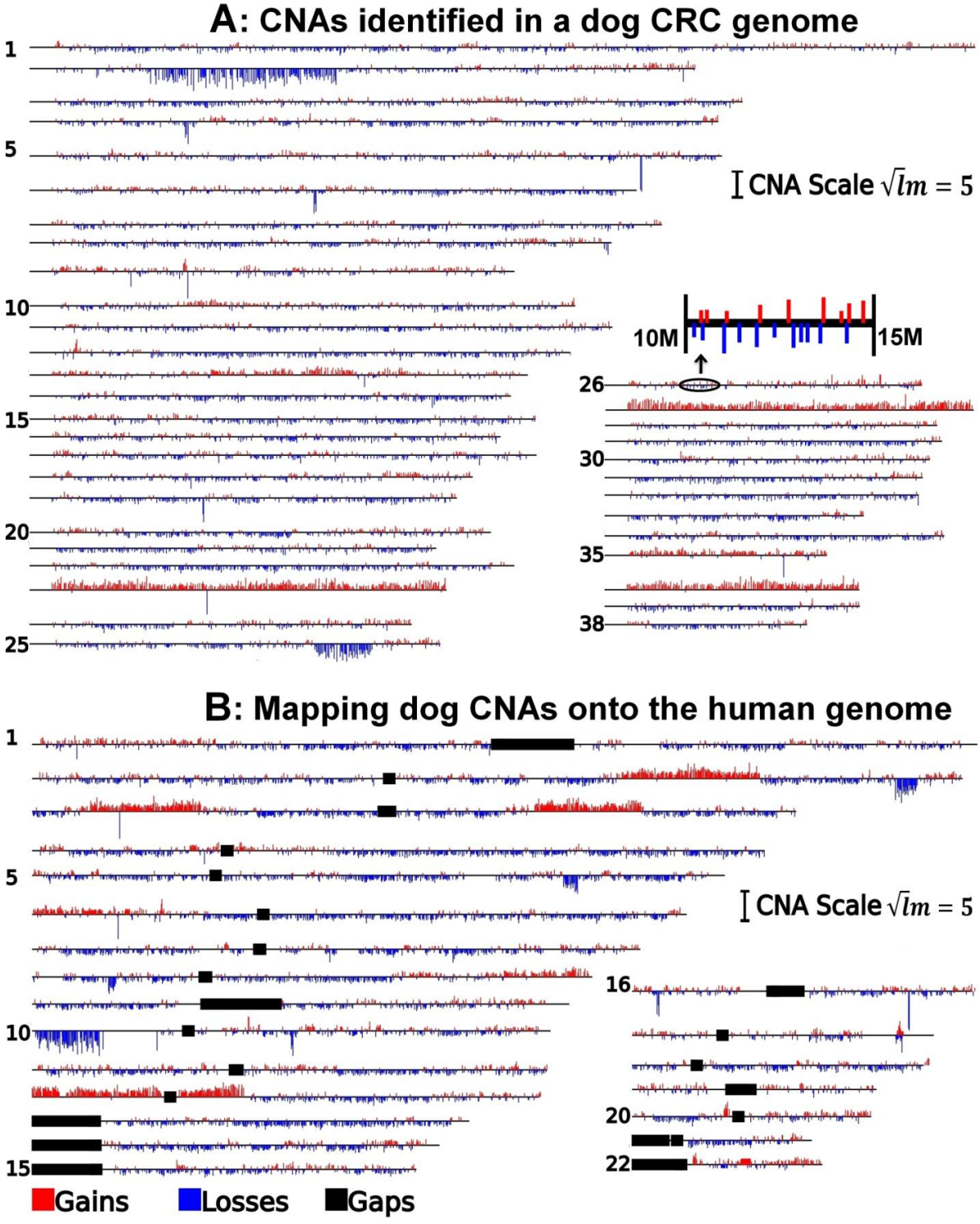

A: CNAs identified in a dog CRC genome
1
5
10
15
20
25
CNA Scale $\sqrt{lm} = 5$
10M
15M
26
30
35
38
B: Mapping dog CNAs onto the human genome
1
5
10
15
16
20
22
CNA Scale $\sqrt{lm} = 5$
Gains
Losses
Gaps

Figure 2.3

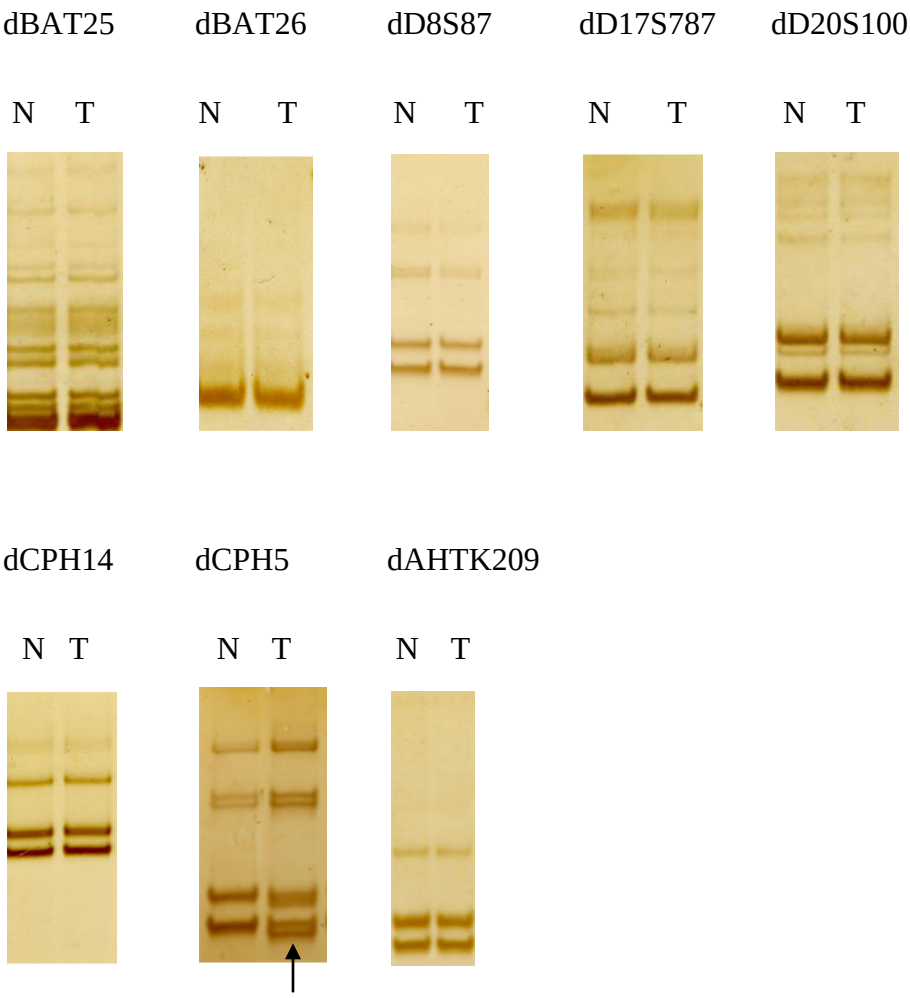

dBAT25
dBAT26
dD8S87
dD17S787
dD20S100
N T
N T
N T
N T
N T
dCPH14
dCPH5
dAHTK209
N T
N T
N T

Figure 2.4

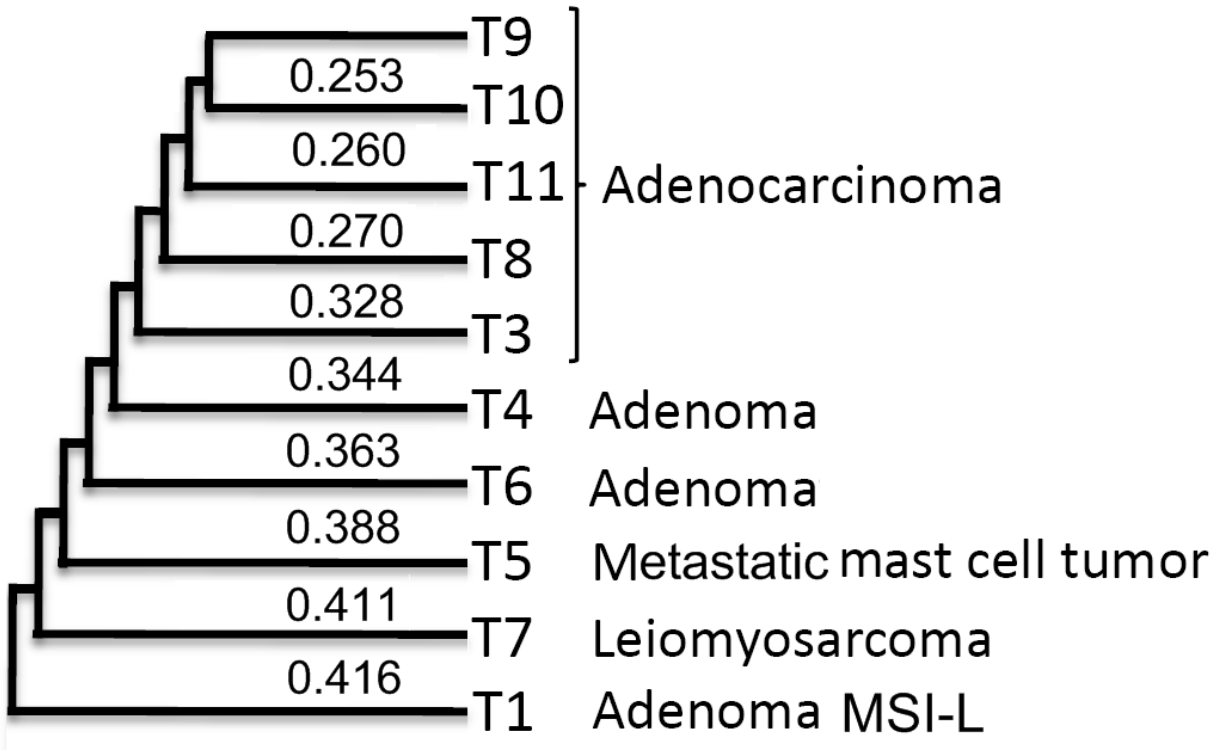


Figure 2.5a

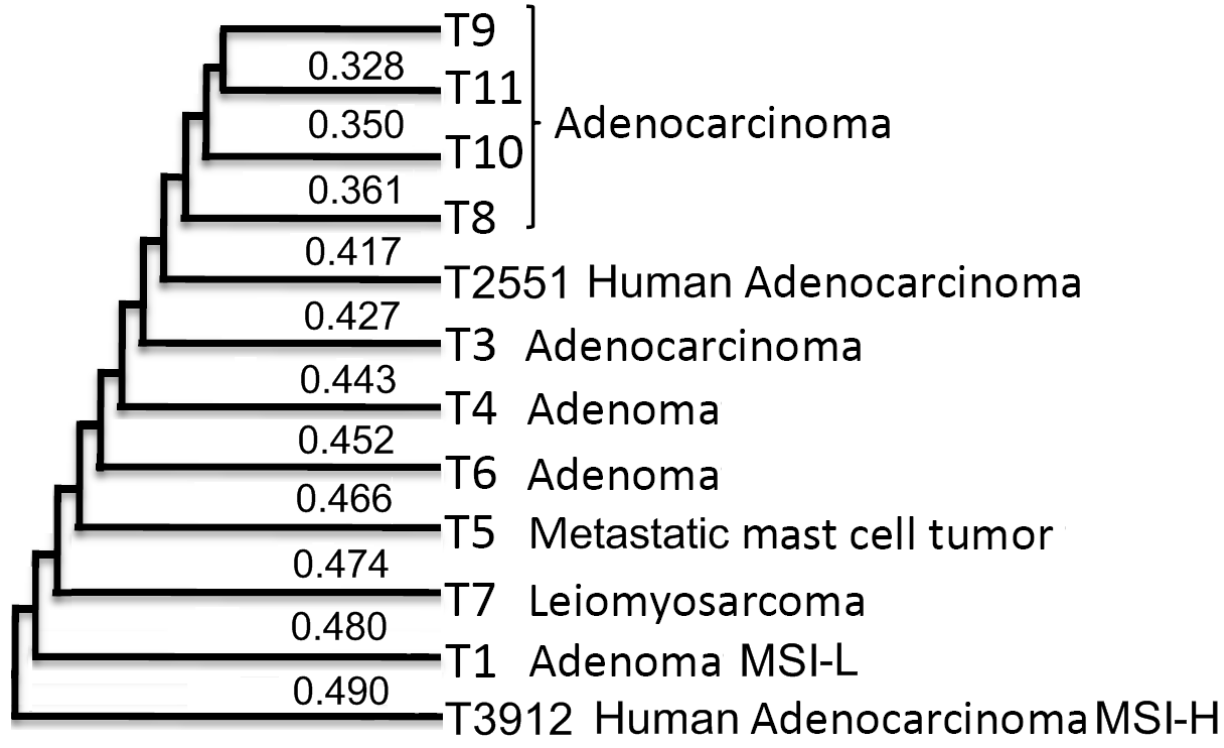


Figure 2.5b

Table 2.1 CNAs identified in 10 dog colorectal tumors[1]

Dog tumors[2]	Total CNA no.	Avg. probe no.	Avg. log_2-ratio	Avg. CNA size, kb	Total CNA size on dog genome, Mb	Total CNA size on human genome, Mb	Total genes
T1, Large Adenoma, MSI-low	2391	11	.188	61	146	151	2136
T4, Adenoma	5708	9	.201	50	285	315	3783
T6, Adenoma	5880	8	.205	43	253	277	3555
T3, Adenocarcinoma	7022	9	.211	50	349	382	4305
T8, Adenocarcinoma	9998	10	.234	54	545	599	6577
T9, Adenocarcinoma	9389	11	.236	58	550	601	6739
T10, Adenocarcinoma	9588	10	.238	54	514	559	6347
T11, Adenocarcinoma	9107	11	.228	60	551	609	5715
T5, Metastasic mast cell tumor	4974	8	.193	40	197	218	3066
T7, Leiomyosarcoma	3486	8	.191	39	138	152	2198

[1]Chromosomes X, Y, M, and UN were not included. [2]Adenomas (T1, T4 & T6) are early-stage tumors, with T1 being more advanced than T4 & T6 and exhibiting a weak MSI phenotype. Adenocarcinomas (T3 & T8-T11) are late-stage tumors. Both adenomas and adenocarcinomas originated from colon epithelial cells. A metastatic mast cell tumor (T5) is a late-stage tumor originated from mast cells, and leiomyosarcoma (T7) is a late-stage tumor originated from smooth muscle cells. [3]The probability of the occurrence of each CNA was calculated as described in **Materials and Methods**; the average of these probabilities for all CNAs found in a tumor is shown here.

Table 2.2a Observed and predicted[1] numbers of disrupted genes that participate in known human CRC pathways[2]

Tumor[3]	Observed#	Predicted#
T1	82	68
T4	144	121
T6	124	114
T3	146	138
T8	226	211
T9	265	216
T10	242	204
T11	228	183
T5	116	98
T7	86	70
T2551	157	146
T3912	30	27

[1]The predicted gene numbers were calculated by assuming that overlap between disrupted genes of each tumor and the human CRC pathway gene members (647 total) is completely random.

[2]The one-tail t-test indicates that observed gene numbers are significantly more than predicted values ($p = 0.0004$).

[3]T1-T11 are dog tumors (see Table 2.1 for more detailed description), whereas T2551 is a human MSI-none adenocarcinoma, and T3912 is a human MSI-high adenocarcinoma.

Table 2.2b Recurrently disrupted dog homologues of genes participating in known human CRC pathways

Pathways	Genes[1]
Apoptosis	*BCL2, FADD, PIK3R3*
Cell cycle	*ABL1, ANAPC10, MAD1L1, MCM5, PRKDC*
MAPK signaling	*CACNA1C, CACNA1D, CACNA1I, CACNA2D3, CACNA2D4, CACNB2, FGF10, FGF12, FGF14, MAP2K5, MAP3K7IP1, MAP3K7IP2, MECOM, NTRK2, PDGFRA, PPM1B, RAPGEF2, RASGRF1, RPS6KA2, RPS6KA5, STK3, TAOK1, ZAK*
P53 signaling	*APAF1, CCND1, MDM4, PPM1D, PTEN, TP73*
TGFβ signaling	*ACVR1, ACVR2A, BMP6, BMPR1B, GDF6, LTBP1*
Wnt signaling	*CTBP2, DKK2, MAPK10, NKD1, NLK, PLCB1, PLCB2, PLCB4, PPP2R5E*

[1]Shown here are significantly disrupted gene with the desired FDR setting to 0.2 (the cutoff p-value was 0.013), following a strategy for recurrent event identification described by Beroukhim et al. (2007).

Table 2.3 Comparison of CNAs between human and dog CRCs

Shared CNAs	Human unique CNAs	Dog unique CNAs
1q31.2-32.3, 1q42.12, 1q42.2-42.3 2q24.1-32.3, 2q34-35 3q24-25.32, 3q26.2-26.32, 3q27.1 4p16.2-15.33, 4p15.2, 4p13-12, 4q12-13.1, 4q21.2, 4q22.1-31.1, 4q31.22-35.1 5p15.3, 5p15.1-14.2, 5p13.3-13.2, 5q23.3 6p25.3-22.1, 6p21.32-21.31, 6p21.1-12.3, 6p12.1, 6q11.2-14.1, 6q15-16.3, 6q22.2-22.31, 6q22.33-23.1, 6q23.3-24.3, 6q25.2-25.3, 6q27 7p22.1, 7p21.2-21.1, 7p15.2-15.1, 7p14.2, 7p12.2, 7q21.13, 7q21.3-31.31, 7q31.33, 7q32.3, 7q36.1, 7q36.3 8p23.3, 8p22-21.2, 8p12-11.21, 8q11.22, 8q12.1, 8q12.3, 8q13.2-13.3, 8q21.12-22.2, 8q23.1-24.11, 8q24.13, 8q24.22-24.3 9p23-22.3, 9p22.1-21.1 11p15.5-14.2, 11p13, 11q22.1-22.2, 11q23.2-23.3 12p13.32-11.21, 12q15-21.1, 12q21.3 13q12.2, 13q13, 13q14.2-14.3, 13q21.2-22.2, 13q31.1-33.2 14q12-22.3, 14q23.2, 14q24.1, 14q24.3-32.33 17p13, 17q11.2-21.2, 17q21.33-22, 17q23.2 18p11.31-11.22, 18q12.3-21.2, 18q22.2, 18q23 19p13.3-13.13, 19p13.11, 19q13.12-13.2, 19q13.32 20p12.1, 20q11.22-13.32	1q24.1-31.1, 1q41-42.11, 1q42.13, 1q43-44 3q25.33-26.1, 3q26.33, 3q27.2-27.3 4p16.3, 4p15.32-15.31, 4p15.1-14, 4q13.2-21.1, 4q21.3, 4q31.21 5p15.2, 5p14.1, 5p13.1-12, 5q23.1-23.2 6p21.33, 6p21.2, 6p12.2, 6p11.2-11.1, 6q11.1, 6q14.2-14.3, 6q21-22.1, 6q22.32, 6q23.2, 6q25.1, 6q26 7p22.3-22.2, 7p21.3, 7p15.3, 7p14.3, 7p14.1-12.3, 7p12.1-11.1, 7q11.1-21.12, 7q21.2, 7q31.32, 7q32.1-32.2, 7q33-35, 7q36.2 8p23.2-23.1, 8p21.1, 8q11.1-11.21, 8q11.23, 8q12.2, 8q13.1, 8q21.11, 8q22.3, 8q24.12, 8q24.21 9p24, 9p22.2 11p14.1, 11p12-11.2, 11q21, 11q22.3-23.1 12p13.33, 12q14.3, 12q21.2 13q11-12.13, 13q12.3, 13q14.11-14.13, 13q21.1, 13q22.3, 13q33.3-34 14q23.1, 14q24.2 17p12-11.2, 17q21.31-21.32, 17q23.1, 17q23.3-25.3 18p11.32, 18p11.21-11.1, 18q11.1-12.2, 18q21.31-22.1, 18q22.3 19p13.12, 19p12-11, 19q11-13.11, 19q13.31, 19q13.33-13.42 20p11.21-11.1	1p36.32-36.12, 1p35.2-32.1, 1p31.1, 1p21.2-13.3, 1p13.1-12, 1q21.2-23.1 2p25.2-16.1, 2p14-13.2, 2p12, 2q12.3, 2q14.1-14.3, 2q21.3-23.3, 2q33.2, 2q36.1-37.3 3p26.3, 3p26.1-24.2, 3p23-21.33, 3p21.2-21.1, 3p14.2, 3p12.3-12.2, 3q12.1-13.12, 3q13.2-21.1, 3q21.3-23, 3q28-29 5q12.1-12.2, 5q13.3-14.2, 5q15-21.2, 5q22.2, 5q31.1-33.1, 5q34-35.2 9p13.2, 9q21.12, 9q21.31-21.32, 9q22.31, 9q31.1-31.2, 9q32-34.3 10p15.2-14, 10p12.33-12.32, 10q21.3-23.2, 10q23.32, 10q24.1-24.31, 10q24.33-25.1, 10q25.3-26.2 11q12.1-14.1, 11q24.1-24.2 12q12, 12q13.12-13.13, 12q13.3-14.2, 12q23.3-24.31 14q11.2, 14q23.3 15q13.3, 15q15, 15q21.3-22.1, 15q22.31, 15q22.33-24.2, 15q25.1, 15q25.3-26.1 16p13.3-16p13.2, 16p12.2-12.1, 16q12.1, 16q13-23.2, 16q24 20q13.33 21q21.1-22.12, 21q22.3 22q11.21, 22q11.23-13.31, 22q13.33

REFERENCES

Albertson DG, Collins C, McCormick F, Gray JW. (2003) Chromosome aberrations in solid tumors. Nat Genet. 34:369-76.

Argyle DJ. Prostate cancer in dogs and men: A unique opportunity to study the disease. Vet J. 2009;180:137-8.

Argyle DJ. The benefits of comparative medicine -- a hundred years to come to our senses. Vet J. 2005;170(2):147-8.

Aebersold et al. Report on EU-USA workshop: how systems biology can advance cancer research (27 October 2008). Mol Oncol. 2009;3:9-17.

Bailey JA, et al. (2002) Recent segmental duplications in the human genome. Science 297, 1003-7.

Bailey et al. Hotspots of mammalian chromosomal evolution. Genome Biology (2004).

Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple-testing. Journal of the Royal Statistical Society. 57(1) 289-300.

Beroukhim R, et al. 2007. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc Natl Acad Sci USA 104: 20007-20012

Boland CR, Thibodeau SN, Hamilton SR, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998;58:5248—5257.

Camps J, Armengol G, del Rey J, Lozano JJ, Vauhkonen H, Prat E, Egozcue J, Sumoy L, Knuutila S, Miró R. 2006. Genome-wide differences between microsatellite stable and unstable colorectal tumors. Carcinogenesis. 27(3):419-2

Camps J, Grade M, Nguyen QT, Hörmann P, Becker S, Hummon AB, Rodriguez V, Chandrasekharappa S, Chen Y, Difilippantonio MJ, Becker H, Ghadimi BM, Ried T. Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. Cancer Res. 2008 Mar 1;68(5):1284-95.

Cormen T.H, et al. (2001) Introduction to Algorithms. 2nd Edition. MIT Press and McGraw-Hill.

Deng L, et al. An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Hum Mutat. 2008 Oct;29(10):1209-16.

Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, Bodmer W, Tomlinson IP, Carter NP. 2004. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer Res. 64(14):4817-25.

Gallahan D, Callahan R. The mouse mammary tumor associated gene INT3 is a unique member of the NOTCH gene family (NOTCH4). Oncogene. 1997;14:1883-90.

Grady WM. Genomic instability and colon cancer. Cancer Metastasis Rev. 2004 Jan-Jun;23(1-2):11-27. Review.

Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446:153-8.

Hampton OA, et al. 2009. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res. 19(2):167-77.

Hanahan & Weinberg. The Hallmarks of Cancer. Cell, Vol. 100, 57–70, 2000.

Hansen K, Khanna C. Spontaneous and genetically engineered animal models: use in preclinical cancer drug development. European Journal of Cancer 40: 858-880, 2004.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6.

Haverty PM, Fridlyand J, Li L, Getz G, Beroukhim R, Lohr S, Wu TD, Cavet G, Zhang Z, Chant J. High-resolution genomic and expression analyses of copy number alterations in breast tumors. Genes Chromosomes Cancer 2008;47:530–42.

Hinoi T, Akyol A, Theisen BK, Ferguson DO, Greenson JK, Williams BO, Cho KR, Fearon ER. Mouse model of colonic adenoma-carcinoma progression based on somatic Apc inactivation. Cancer Res. 2007 Oct 15;67(20):9721-30.

Ji X, Zhao S (2008). DA and Xiao – Two Composite and Giant LTR-retrotransposon like Structures Identified on the Human Genome. Genomics, 91(3):249-258.

Jones S, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008; 321:1801-6.

Kaiser S, et al. Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. Genome Biol. 2007;8(7):R131.

Khanna C, et al. The dog as a cancer model. Nat Biotechnol. 2006;24:1065-6.

Kidd JM, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 453 2008: 56-64

Kinzler K. W., Vogelstein B. Lessons from hereditary colorectal cancer. Cell, 87: 159-170, 1996.

Knofczynski GT. Sample Sizes when Using Multiple Linear Regression for Prediction *Educational and Psychological Measurement*, v68 n3 p431-442. 2008

Knutsen T, Gobu V, Knaus R, Padilla-Nash H, Augustus M, Strausberg RL, Kirsch IR, Sirotkin K, Ried T. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. Genes Chromosomes Cancer. 2005;44:52-64.

Kruskal, J.B. (1956) On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proceedings of the American Mathematical Society, 7, 48–50.

Lengauer C, Kinzler KW, Vogelstein B. 1997. Genetic instability in colorectal cancers. Nature. 386:623-7.

Leroy BE, Northrup N. Prostate cancer in dogs: Comparative and clinical aspects. Vet J. 2009;180:149-62.

Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803-19 (2005).

Liu, Jun, et al. (2006) Distance-based clustering of CGH data. Bioinformatics, 22, 1971-1978.

Maser RS, et al. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. Nature. 2007;447:966-71.

McNiel EA, Griffin KL, Mellett AM, Madrill NJ, Mickelson JR. Microsatellite instability in canine mammary gland tumors. J Vet Intern Med. 2007 Sep-Oct;21(5):1034-40.

Michell, A.R., 2004. Comparative Clinical Science: the Medicine of the Future. The Veterinary Journal, doi: 10.1016/j.tvjl. 2004.06.004.

Moser AR, Mattes EM, Dove WF, Lindstrom MJ, Haag JD, Gould MN. 1993. ApcMin, a mutation in the murine Apc gene, predisposes to mammary carcinomas and focal alveolar hyperplasias. PNAS 90:8977-81.

Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 309:613-7.

Nakao K, et al. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. Carcinogenesis. 2004 25(8):1345-57.

Paoloni M, Khanna C. Translation of new cancer treatments from pet dogs to humans. Nat Rev Cancer. 2008;8:147-56.

Paoloni MC, Khanna C. Comparative oncology today. Vet Clin North Am Small Anim Pract. 2007;37:1023-32.

Pevzner, P. & Tesler, G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. PNAS 100, 7672-

Prim, R.C. (1957) Shortest connection networks and some generalisations. Bell System Technical Journal, 36, 1389–1401.

Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C. (2003) The significance of unstable chromosomes in colorectal cancer. Nat Rev Cancer. 2003 Sep;3(9):695-701.

Rosol TJ, Tannehill-Gregg SH, LeRoy BE, Mandl S, Contag CH. Animal models of bone metastasis. Cancer 97(3 Suppl): 748-757, 2003.

Sancho, E; Batlle, E; Clevers, H. Signaling pathways in intestinal development and cancer. Annu Rev Cell Dev Biol. 2004;20:695–723.

Steinbach,M. et al. (2000) A comparison of document clustering techniques. KDD Workshop on Text Mining, Boston, MA, USA.

Turner DJ, et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nature Genetics 40: 90-95, 2007.

Toft N. J., Arends M. J. DNA mismatch repair and colorectal cancer. J. Pathol., 185: 123-129, 1998.

Vail DM, MacEwen EG. Spontaneously occurring tumors of companion animals as models for human cancer. Cancer Investigation, 18(8): 781-792, 2000.

Yang L, Hamilton SR, Sood A, Kuwai T, Ellis L, Sanguino A, Lopez-Berestein G, Boyd DD. The previously undescribed ZKSCAN3 (ZNF306) is a novel "driver" of colorectal cancer progression. Cancer Res. 68:4321-30. 2008

Zhao, S, et al. (2004) Human, Mouse and Rat Genome Large Scale Rearrangements: Stability Versus Speciation. Genome Res. 14:1851-60.

CHAPTER 3

CANCER DRIVER-PASSENGER DISTINCTION VIA SPORADIC HUMAN AND DOG CANCER COMPARISON: A PROOF OF PRINCIPLE STUDY WITH COLORECTAL CANCER[1]

[1]Tang J, Li Y, Lyon K, Camps J, Ried T, Zhao S. Submitted to *Genome Research,* 03/15/2012.

ABSTRACT

We report herein a proof of principle study illustrating a novel human-dog comparison strategy that addresses a central aim of cancer research: cancer driver–passenger distinction. We previously demonstrated that sporadic canine colorectal cancers (CRCs) share similar molecular pathogenesis mechanisms as their human counterparts. In this study, we compared the genome-wide copy number abnormalities between 29 human- and 10 canine sporadic CRCs. This led to the identification of 73 driver candidate genes (DCGs), altered in both species and with 27 from the whole genome and 46 from human-dog genomic rearrangement breakpoint (GRB) regions, as well as 38 passenger candidate genes (PCGs), altered in humans only and located in GRB regions. We noted that DCGs significantly differ from PCGs in every analysis conducted. Importantly, while PCGs are not enriched in any specific functions, DCGs possess significantly enhanced functionality in establishing and maintaining epithelial cell apicobasal polarity. Meanwhile, many DCGs also participate in processes that regulate cell proliferation and death. These observations indicate that, in sporadic CRCs of both species, cell polarity genes not only play a critical role in preventing cancer cell invasion and spreading, but also likely serve as tumor suppressors by modulating cell growth. This pilot study validates our novel strategy, expansion of which would make more driver-passenger distinctions, and address key questions regarding the relationship between cancer pathogenesis and epithelial cell polarity control.

Running Title: Driver-passenger distinction by human-dog comparison

Key words: cancer driver-passenger distinction; sporadic human-dog CRC comparison; epithelial cell apicobasal polarity; new potential colorectal tumor suppressor genes

INTRODUCTION

A central aim of cancer research has been to identify a handful of cancer-causative alterations (drivers) from hundreds to thousands of abnormal changes found in a cancer genome (Haber and Settleman 2007; Stratton et al. 2009; Ledford 2010). Discoveries of such drivers in the past have contributed to the understanding of cancer etiology and yielded prognostic markers and therapeutic intervention targets such as BCR-ABL, leading to the development of the spectacularly successful anti-leukemia drug imatinib (Gleevec) (Sawyers 1999). As high throughput technologies such as next-generation sequencing and high density array become routinely available, hundreds of thousands of cancer genomes will be characterized and millions of cancer-related somatic mutations will be uncovered (Haber and Settleman 2007; Pasche and Myers 2009; Stratton et al. 2009; Hudson et al. 2010; Ledford 2010; Gibbs 2011). Hence, the need to efficiently distinguish drivers from passengers (incidental changes or changes occurring as a consequence of cancer) becomes increasingly pressing, and has become one of the National Cancer Institute (NCI)'s Provocative Questions that need to be urgently addressed (provocativequestions.nci.nih.gov).

Drivers are typically more recurrent than passengers, and researchers have been using strategies such as increasing sample size and studying early stage cancers to distinguish them (Futreal et al. 2004; Haber and Settleman 2007; Stratton et al. 2009; Hudson et al. 2010; Ledford 2010). Differing from these traditional approaches that study human cancers only, we have developed a novel human-dog comparative genomics and oncology strategy for cancer driver-passenger distinction, rationalized as follows.

The dog has become an increasingly important model for studying human physiology and diseases (Cyranoski 2010), and sporadic canine cancer represents one of the best cancer models (Rowell et al. 2011). First, these cancers are naturally-occurring and heterogeneous, unlike most genetically-modified or xenograft rodent models, and significantly, dogs share the same environment as humans and hence are exposed to the same carcinogens. Indeed, our studies have revealed that canine sporadic colorectal cancers (CRCs) share similar molecular pathogenesis mechanisms as their human counterparts (Youmans et al. ; Tang et al. 2010). Furthermore, the dog genome has been sequenced to a >7X coverage and a relatively accurate version of its genome sequence assembly is available (Lindblad-Toh et al. 2005), unlike another companion animal, the cat (Pontius et al. 2007).

Critically, the dog genome is rearranged when compared to the human genome with over 300 inversions and translocations identified (Lindblad-Toh et al. 2005; Ji and Zhao 2008), resulting in thousands of genes that are clustered in the human genome to be dispersedly located in the dog genome. We utilize these orthologous genes' different genomic locations between the two species for driver-passenger distinction. This fundamentally enables our cross-species comparison strategy to better distinguish drivers from passengers than human cancer-only approaches for cancers with large genomic deletions/amplifications, to which the vast majority of epithelial cancers belong, as illustrated by our pilot study described below.

RESULTS AND DISCUSSION

CRC driver-passenger distinction via human-dog comparison

To illustrate this novel human-dog comparison strategy, we conducted a proof of principle study on CRC, one of the best understood cancers for studying cancer initiation and progression (Fishel et al. 1993; Kinzler and Vogelstein 1996; Sjoblom et al. 2006) and for which we have previously demonstrated numerous human-dog molecular similarities (Youmans et al. ; Tang et al. 2010). We compared the genome-wide copy number abnormalities (CNAs) between 29 human sporadic CRCs (Camps et al. 2008) and 10 canine sporadic CRCs (Tang et al. 2010), with the vast majority being CIN (chromosome instability) tumors and harboring large genomic amplifications and deletions. This analysis first led to the discovery of 27 genes, with each having at least one of its exons significantly altered in both species (Figure 3.1), genome-wide [sex chromosomes excluded because the canine Y chromosome is not yet sequenced (Lindblad-Toh et al. 2005), and the extensive sequence duplication between X and Y chromosomes complicates the analysis (Skaletsky et al. 2003)]. These 27 genes, which include the best known colorectal tumor suppressor *APC* (Kinzler and Vogelstein 1996), are referred to as the first driver candidate genes (DCGs) or 1stDCGs hereafter. Next, we focused on the human genomic sites that harbor the human-dog genomic rearrangement breakpoint (GRB) (a total of 324 GRBs were detected for the autosomes) to identify potential driver-passenger pairs (Figure 3.1), based on the following criteria. First, the greater GRB region, which includes the GRB itself and two immediate flanking regions with each containing at least one gene, is altered in at least one human tumor. In addition, at least one canine orthologous gene at one side of the GRB is altered in at least

one dog tumor (genes changed in both species were deemed drivers), while none of the canine orthologous genes at the other side of the GRB is disrupted in any of the dog tumors (these genes were considered as passengers) (Figure 3.1). This allowed the identification of additional 46 DCGs (referred to as 2ndDCGs hereafter) and 38 passenger candidate genes (PCGs) from 25 GRB sites meeting the specified conditions, with most genes in both DCG and PCG groups recurrently altered in the human CRCs.

Known cancer drivers found among DCGs but not among PCGs

Besides *APC*, both 1stDCGs and 2ndDCGs contain known cancer genes with many already being targeted for therapeutic intervention, including the DNA damage-inducible cell-cycle regulator *GADD45A* (Rosemary Siafakas and Richardson 2009; Reinhardt et al. 2010), the constitutive glucose transporter *SLC2A1* (Amann and Hellerbrand 2009), the signaling molecule A3 adenosine receptor *ADORA3* (Jacobson et al. 2009), the chromatin modifier *EZH2* (Futreal et al. 2004), the receptor tyrosine phosphatase *PTPRD* (Veeriah et al. 2009), and the cancer biomarker *MUC16* (Bouanene and Miled 2010). The PCGs, however, do not contain any such genes. Below, we will describe more analyses that further examine their differences.

DCGs significantly differ from PCGs in cancer relevance, biological functions, and drug response examined

We first took advantage of existing databases and published literature to assess the cancer relevance and known biological functions of our candidate genes. We determined the presence/absence of these genes in well-known cancer databases

including Cancer Gene Census (a catalog of genes of which mutations are causally implicated in cancer) (Futreal et al. 2004), KEGG cancer pathways, COSMIC (a catalog of somatic mutations of cancers) (Forbes et al. 2011), and Cancer Gene Index (a database of associations among genes, cancers, and drug compounds developed at the NCI). Then, we searched the Mouse Genome Informatics database for phenotypic changes in existing gene-knockout mouse models. It has been reported that cancer genes interact with more partners than a typical human gene (Lin et al. 2007); we thus also determined the predicted protein-protein interactions and DNA-protein interactions (which imply how well a gene is regulated in its transcription) of each gene. Finally, we counted the number of publications of each gene in PubMed as a measure of its functional importance, with the assumption that already well-studied genes may play critical roles in biological processes. All the information was catalogued, on which we performed Hotelling's T-squared tests. As shown in Figure 3.2, while the difference between the two DCG groups is insignificant, DCGs indeed differ significantly from PCGs in the examined known cancer relevance and biological functions. This supports that our cross-species comparison approach is valid for driver-passenger distinction.

To further examine the DCG-PCG difference and similarity, we treated HCT116, a near diploid CRC line, with the anticancer drug Taxol (Paclitaxel). Taxol has been shown to stabilize microtubules and prevent them from disassembly, blocking the progression of cell cycle and triggering cell apoptosis (Long and Fairchild 1994; Yvon et al. 1999; Jordan and Wilson 2004). For PCGs, overall, we did not observe any significant changes in gene expression at different Taxol concentrations (10 nM, 20 nM,

and 80 nM) and treatment times (24 hour and 48 hour) (Figure 3.2). However, for both 1stDCGs and 2ndDCGs, we noted an initial small expression decrease after 24 hour Taxol treatment for most genes, which made the overall expression decrease significant, followed by a larger expression increase after 48 hour making the expression level even higher than before the drug treatment for many genes (Figure 3.2). Two DCGs, *GADD45A* and *C1orf63,* are especially noteworthy, having their expression level increasing along with the Taxol concentration and incubation time and reaching 18-fold for *C1orf63* and 14-fold for *GADD45A* at 80nM for 48 hour (Figure 3.2). *GADD45A* is known to be upregulated by stress and drug therapy, and is required at the G_2/M checkpoint and arrests cell cycle progression (Sheikh et al. 2000; Rosemary Siafakas and Richardson 2009; Reinhardt et al. 2010). The function of *C1orf63* is unknown currently; its similar Taxol response profile as *GADD45A* (Figure 3.2) raises the possibility that *C1orf63* may possess similar functions as *GADD45A*. In summary, while much more work is clearly needed to understand the mechanisms of the observed drug response, the study demonstrated a clear difference between DCGs and PCGs in their overall response to Taxol treatment in HCT116 cells, supporting that our approach is valid.

1stDCGs enriched in cell adhesion/motility functions, 2ndDCG enriched in vesicle-trafficking and ion-transport functions, whereas PCGs not enriched in any specific functions

We applied the DAVID Gene Functional Classification Tool (Huang da et al. 2009) on the candidate genes and observed notably more functional groups being over-represented in DCGs than in PCGs. These include cell adhesion (p=0.02), adherent

junctions (p=0.04), cell motility (p=0.05), and cell morphogenesis (p=0.05) for 1stDCGs, as well as vesicle-trafficking (p=0.03) for 2ndDCGs. They make up a subset of the functional groups reported by a previous human CRC sequencing study (Sjoblom et al. 2006). For PCGs, however, no any significant enrichment was found, with C2H2-zinc finger (ZNF) genes being the most enriched group ($p = 0.14$).

Because 2ndDCGs and PCGs were identified from the 324 GRB sites rather than genome-wide (Figure 3.1), we performed the same enrichment analysis using the 4,401 genes encoded in 1Mb sequences surrounding each GRB (0.5Mb each of upstream and downstream of the GRB center) as the background, instead of the entire autosome gene set as above. The analysis revealed ion homeostasis/transport and vesicle-trafficking remaining significant ($p \leq 0.05$) for 2ndDCGs as above. However, for PCGs, still no significant enrichment was observed, and C2H2-ZNFs (p=0.23) were replaced by nuclear lumen-associated genes as the most enriched, which stay nevertheless insignificant (p=0.17).

GRB sites enriched in C2H2-ZNFs and G-protein coupled receptors (GPCRs)

The following analysis explained why up to five C2H2-ZNFs are among PCGs. Compared to the whole genome (sex chromosomes excluded), the 324 GBR sites, which were used to identify PCGs and 2ndDCGs (Figure 3.1), are significantly enriched with C2H2-ZNF genes ($p < 0.0001$). However, when comparing the 25 GBR sites, where PCGs and 2ndDCGs locate, against the entire 324 GBR set, the difference is insignificant ($p > 0.3$). Hence, the GRB sites harbor significantly more C2H2-ZNFs than other

genomic regions, and the five C2H2-ZNFs in PCGs were pulled out by random chance. Similar conclusions were reached for GPCRs and olfactory receptors (ORs) (four GPCRs, of which three are ORs, are among PCGs, while only one is among DCGs). C2H2-ZNFs and GPCRs are the only gene families of which multiple members were found among PCGs.

Other DCG-PCG differences in functional groups

Besides those identified by the DAVID software (Huang da et al. 2009) as described above, we also noted several other functional groups being more prominent in DCGs than in PCGs. These include: 1) small GTPase signaling genes, with five found for DCGs but only one identified for PCGs, contrary to the GPCRs described above; 2) ubiquitin-dependent protein degradation genes (six for DCGs versus none for PCGs; 3) redox genes (five for DCGs versus none for PCGs); 4) calcium-binding/dependent genes (10 for DCGs versus one for PCGs); and 5) developmental genes (13 for DCGs versus three for PCGs).

The majority of DCGs function in cell polarity, while PCGs do not

Colon epithelial cells are asymmetric, exhibiting apical and basal polarity (Figure 3.3). This allows them to form permeability barriers between two compartments in the body and to vectorially transport ions and solutes between the compartments. According to Nelson and others (Nelson 2003), development of epithelial cell polarity requires cell-cell and cell-substratum adhesions achieved by genes functioning in various adhesion complexes, cell junctions, and cytoskeleton organization. In addition, the polarized

plasma membrane, which is divided into functionally and structurally distinct apical and basolateral domains, is established and maintained by polarized intracellular protein sorting, trafficking and targeting involving a variety of genes and organelles such as trans-Golgi network (TGN) and endosomes.

We found that up to 78% of 1stDCGs and 72% of 2ndDCGs, but only 26% of PCGs, could participate in these processes of establishing and maintaining epithelial cell polarity. Specifically, as many as 56% of 1stDCGs (15 genes out of 27 total, with 12 certain and three likely) and 39% of 2ndDCGs (18 genes out of 46 total, with 13 certain and five likely), but only 13% of PCGs (five genes out of 38 total, with three certain and two likely), possess functionality in cell-cell or cell-substratum adhesions, as listed in Figure 3.3. For polarized ion-transport and trafficking, we noted up to 33% of 2ndDCGs (15 genes with 13 certain and two likely) and 22% of 1stDCGs (six genes with four certain and two likely), but only 13% of PCGs (five genes) having or likely having such functions (Figure 3.3).

Significantly, a substantial portion of DCGs are already known to be located in different portions of the polarized plasma membrane, as listed in Figure 3.3; however, we did not find such information for any of the PCGs. Furthermore, a total of eight 1stDCGs and six 2ndDCGs are already-known epithelial cell polarity genes (Figure 3.3). Many function in cell adhesion, including protocaherin *PCDH9*; the multifunctional CRC tumor suppressor *APC;* membrane-associated guanylate kinase (MAGUK) *MPP7* and *DLGAP2* (*DLG*-associated protein 2) (Bryant and Mostov 2008); basement membrane genes

LAMA1 and *FREM2*; cytoskeleton genes *ACTC1* and *KIFAP3* (Li and Gundersen 2008); and small GTPase signaling molecules *RAP1A* and *ARHGEF7* (Iden and Collard 2008). Others function in polarized ion and small molecule transporting, including potassium/chloride cotransporter *SLC12A6* (Bachmann et al. 2011), glucose/cation symporter *SLC2A1*, apical membrane-located v-type proton ATPase subunit *ATP6V1B2* (Belleannee et al. 2010), and phosphodiesterase *ENPP1* (Belleannee et al. 2010). Furthermore, 1stDCGs contain three additional potential cell polarity genes: 1) *EFNB2*, a member of Eph/Ephrin signaling pathway which regulates the mesenchymal-epithelial transition (MET)(Barrios et al. 2003); 2) EYA2, a tyrosine phosphatase that dephosphorylates histone H2AX and involves in epithelial-mesenchymal transition (EMT)(Farabaugh et al. 2011); and 3) *MYCBP2*, a *MYC*-binding protein whose functions include cytoskeleton organization (Li and Gundersen 2008). Similarly, 2ndDCGs also harbor five such genes: 1) *AVL9*, an exocytosis gene in yeast (Harsay and Schekman 2007); 2) *CHRNA7*, a cholinergic receptor whose functions include regulating calcium homeostasis (Krais et al. 2011); 3) *EPDR1*, an ependymin-related gene that is involved a cell-substratum adhesion; 4) *MLANA*, a palmitoylated integral membrane protein likely involving in sorting and degradation of melanosome proteins (Levy et al. 2005); and 5) *SFRP4*, a WNT-signaling gene. In summary, as many as 41% of 1stDCGs and 24% of 2ndDCGs are already-known or likely polarity genes. For PCGs, only three (8%) are possibly such genes [*CTGF*, *SEC24C*, *TRIM62* (Lott et al. 2009)].

Deleted (in CRC) DCGs' expressions increase significantly in a cell culture system where cell-cell/cell-ECM (extracellular matrix) adhesions are establishing

The above analysis is corroborated by mRNA expression changes of DCGs and PCGs in an *in vitro* human embryonic stem cell (hESC) differentiation system, hESC line WA09 → *ISL1+* nascent mesoderm (INM), during which cell-cell and cell-ECM adhesions are being built (unpublished data). As summarized in Figure 3.4, we observed a significant expression increase on the whole for DCGs that were predominantly deleted in CRCs, which is especially evident for those from 1stDCGs where the overall expression increased by two-fold. Meanwhile, the expression of PCGs barely changed on average. This result is consistent with the cell adhesion function listed for many of the deleted DCGs in Figure 3.3, and supports that many DCGs participate in cell polarity building whereas PCGs do not.

Polarity genes are cancer driver genes?

Loss of cell polarity is a hallmark of cancers that originate from epithelial cells (Royer and Lu 2011). Although its role as a driver or a passenger of cancer is still under debate at present (Royer and Lu 2011), cell polarity has been proposed to be critical in cancer initiation by controlling asymmetric division of cancer stem cells, as well as in cancer progression by preventing cancer cell invasion and spreading (Lee and Vasioukhin 2008). This has become increasingly appreciated in recent years as more studies demonstrate the importance of the EMT in cancer (Kalluri and Weinberg 2009).

This study also indicates that cell polarity genes are important in CRC pathogenesis. First, as described previously, nearly half of our DCGs function in cell-cell adhesion or cell-substratum adhesion. Disruption of these genes would free tumor cells to invade tissue layers below the basement membrane and spread to other places (Figure 3.3), which indeed occurred in all human CRCs [classified T1, T2, T3 or T4 tumors (Camps et al. 2008)] and most of the canine CRCs [many are invasive adenocarcinomas (Tang et al. 2010)] investigated in this study. Hence, alterations of these cell polarity genes are indeed cancer drivers in this regard. Second, as listed in Figure 3.3, besides the known cancer genes described previously, additional 13 DCGs are recognized to partake in processes that control cell proliferation and death, and two extra (*MYBP2* and *ZZEF1*) participate in cell cycle regulation.

Furthermore, we have successfully demonstrated that siRNA knockdown of *RASA3*, *DENND5A*, *AVL9*, and *NUPL1*, all predominantly deleted in CRCs, indeed promoted cell growth (Figure 3.5). These observations support that the four genes are potential tumor suppressors of CRC. As listed in Figure 3.3, they are also associated with epithelial cell polarity, with *RASA3* (a *RAS* GTPase activator) and *DENND5A* [a GDP-GTP exchange factor for *RAB39*; see (Yoshimura et al. 2010)] both participating in small GTPase signaling, *AVL9* functioning in exocytosis in yeast (Harsay and Schekman 2007), and *NUPL1* being a nucleoporin gene (Figure3.3). Except for *RASA3* which is known to inactivate *RAS* by enhancing its weak intrinsic GTPase activity (Nafisi et al., 2008), no any previous studies have been reported in the literature regarding the role of these genes in cancer.

In summary, 26 out of 73 total (~36%) DCGs function in cell proliferation and death regulation, cell cycle control, and/or DNA repair (Figure 3.3); thus, they are traditional cancer driver genes. For PCGs, however, only three out of 38 total (<10%) are involved in such processes (*ASAH2B, CTGF,* and *SERBP1*), some of which may be false negatives that could be resolved by increasing sample size. Most critically, among the 26 DCGs listed in Figure 3.3 as traditional cancer drivers, 21 of them (81%) possess cell polarity functionality. Hence, they are similar to genes such as *PARD3* and *ASPP2*, which are known to control both cell polarity and cell proliferation (Sottocornola et al. 2010)

CONCLUSION

This pilot project provides a proof of principle study successfully demonstrating that our human-dog comparison strategy for CRC driver-passenger distinction is valid. First, the identified DCGs significantly differ from the identified PCGs in every analysis performed. More importantly, we discovered that, while PCGs are not enriched in any specific functions, DCGs hold significantly enhanced functionalities that are closely associated with epithelial cell polarity establishment and maintenance. Meanwhile, many DCGs participate in processes that regulate cell proliferation and death, and/or cell cycle. Hence, this pilot study indicates that, in sporadic CRCs of both species, genes that establish and maintain epithelial cell polarity not only play a critical role in preventing cancer cell invasion and spreading, but also likely serve as tumor suppressors by modulating cell growth.

Unlike previously published papers that are limited to the mere identification of genes recurrently altered in CRCs [e.g., (Sjoblom et al. 2006)], this human-dog comparison approach allowed us to further classify these genes as either cancer driver candidates or passenger candidates, effectively narrowing down the targets for downstream functional validation. Expanding this pilot study to a larger scale would likely make many more driver-passenger distinctions and address key questions regarding cancer pathogenesis and epithelial cell polarity control.

EXPERIMENTAL PROCEDURES

All analyses were performed based on the canFam version 2.0 and the human genome NCBI build 36.1 (hg18). CNAs in both human (Camps et al. 2008) and dog (Tang et al. 2010) CRCs were identified as described previously (Tang et al. 2010). Amplified/deleted genes were identified using the RefGene annotation for the human, and the xenoRefGene annotation for the dog, both downloaded from the University of California Santa Cruz (UCSC) genome site (www.genome.ucsc.edu). Genes recurrently disrupted in both species were identified with the GISTIC algorithm (Beroukhim et al. 2007). Human/dog genomic synteny and rearrangement data were obtained by integrating our data (Ji and Zhao 2008) and the UCSC hg18-canfam2 net alignment data.

The Cancer Gene Census (Futreal et al. 2004) and COSMIC (Forbes et al. 2011) data v48 were both from www.sanger.ac.uk, while the Cancer Gene Index was from cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index. Other databases used included the KEGG Release 54.0 human cancer pathways (hsa05200-hsa05223) from

www.genome.jp/kegg/, and the MGI phenotype database v4.35 from www.informatics.jax.org. Protein-protein interactions were determined using the I2D database v1.9 at ophid.utoronto.ca/, while DNA-protein interactions were determined by counting the transcription factor binding sites (TFBSs) in DAVID v6.7 (Huang da et al. 2009) (david.abcc.ncifcrf.gov/). Gene functional classification was investigated using DAVID v6.7 (Huang da et al. 2009), the Pfam database (pfam.sanger.ac.uk), and literature reports.

HCT116 cells were cultured following the instruction provided by ATCC. Taxol (Paclitaxel) was purchased from Sigma-Aldrich (Product Number: T7402). The expression levels of 1stDCGs, 2ndDCGs, and PCGs in HCT116 cells (with or with Taxol-treatment at different concentrations), as well as WA09 and INM cells were determined by quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) analyses as described previously (Ji et al. 2010). *GAPDH* was used as the normalization gene. siRNA primers were designed using a Silencer® siRNA Construction Kit Template Design Tool and siRNA Target Finder (Ambion). Then, siRNA and scramble-control RNA for each gene were synthesized with a *Silencer*® siRNA Construction Kit (cat no. AM1620) from Ambion. Transfection was performed with siPORT *NeoFX* Transfection Agent (cat no. AM4510) from Ambion. HeLa cell growth and crystal violet-staining were performed as described (Beck et al. 2010). Images were quantified using ImagJ (rsbweb.nih.gov/ij/) and t-tests were performed between scramble-control and knockdown.

Acknowledgements

This work was funded by the American Cancer Society and the Georgia Cancer Coalition (PI: S. Zhao), as well as NCI P50 CA128613 (PI: Dr. Dong M Shin) and GM085354 (PI: Dr. Stephen Dalton). We would like to thank Ms. Huan Xiong for her work qRT-PCR experiments; Dr. Stephen Dalton for providing WA09 and INM cells; and Drs. J David Puett, Dong M Shin, Claiborne Glover, and Georgia Chen for their help and advice on this study.

FIGURE LEGENDS

Figure 3.1. CRC driver-passenger distinction via human-dog comparison. The top portion illustrates how 1stDCGs (27 total), shown in purple and with at least one of their exons significantly amplified/deleted in both species, were identified genome-wide (sex chromosomes excluded; see the main text for more details). The middle section indicates how 2ndDCGs (46 total), shown in green, and PCGs (38 total), shown in red, were identified from the human-dog GRB sites in the human genome (see the main text for criteria for DCG-PCG distinction for more details). The bottom area shows examples of DCG-PCG pairs identified from two GRB sites in the human genome that harbor a human-dog translocation breakpoint (left) or a human-dog inversion breakpoint (right). Genes in each example are clustered in the human genome, but are either located on different chromosomes (left), or far apart (0.2 Mb for the human versus 16.6Mb for the dog) if on the same chromosome (right), in the dog genome. In addition, genes in each example are all amplified in the human tumors; however, in the dog tumors, only DCGs [*GADD45A* and *SLC2A1*, both of which are already known cancer driver genes (Amann and Hellerbrand 2009; Rosemary Siafakas and Richardson 2009; Reinhardt et al. 2010)] are amplified and PCGs (genes shown in red, none of which has been reported in the literature to be cancer driver genes) remain neither amplified nor deleted.

Figure 3.2. A) DCGs significantly differ from PCGs in cancer relevance and known biological functions examined. Hotelling's T-squared tests were performed with the data described in the text. The numbers shown between the boxes are p-values from the tests.

B) DCGs and PCGs responding differently to Taxol-treatment in HCT116 cells. The treatment of 20 nM Taxol for 24 hour and 48 hour induced HCT116 cell death (Top). The overall gene expression change at 20nM Taxol between different treatment times was indicated by the p-values, obtained from t-tests with "↑" representing significant increase and "↓" indicating significant decrease (middle). The bottom shows the expression change of individual DCGs, including *GADD45A* and *C1orf63* with similar drug-response profiles, as well as those having a slight expression decrease at 24 hour followed by a larger increase or a complete recovery at 48 hour. Similar gene expression trends were observed for Taxol treatment at 10nM and 80nM.

Figure 3.3. DCGs (but not PCGs) are enriched in functions in epithelial cell polarity establishment and maintenance, and many DGGs participate in cell proliferation/death and/or cell cycle control. The top left images indicate that loss of epithelial cell polarity occurs in both adenomas (tumor cells not yet penetrating the basement membrane) and adenocarcinomas (tumor cell already invading tissue layers below the epithelium). The genes shown at the right of the images are 1stDCGs (purple), 2ndDCGs (green), or PCGs (red) already known or likely functioning in epithelial cell polarity, based on gene ontology (GO) terms, INTERPRO domains, and other information provided by DAVID (Huang da et al. 2009), as well as literature reports [see the main text and (Argueso et al. 2009)]. The genes with "?" were classified based on protein domain matches only. The genes shown below the images are DCGs or PCGs participating in cell proliferation/death and/or cell cycle regulations, based on annotation

by DAVID (Huang da et al. 2009), published literature (see the main text), and experiments (Figure 3.5).

Figure 3.4. Expression changes of DCGs and PCGs in an *in vitro* human embryonic stem cell (hESC) differentiation system, during which cell-cell/cell-ECM adhesions are establishing. The expression level of individual genes of 1stDCGs, 2ndDCGs, and PCGs in hESC line WA09 and its differentiated *ISL1*+ nascent mesoderm (INM) was determined by qRT-PCR. T-tests were then performed to determine the significance of the expression difference for the genes of each group, which were further separated depending upon if a gene was predominantly amplified (indicated by "Amplification" in the image) or deleted (indicated by "Deletion") in CRCs examined, between the two cell types.

Figure 3.5. siRNA knockdown of deleted (in CRCs) DCGs promotes cell growth. **A**) The four DCGs (*RASA3, DENND5A, AVL9,* and *NUPL1)* were predominantly deleted in CRCs examined. In HCT116 cells, these genes are expressed and siRNA knockdown was performed, as shown by the bar graphs on the top where the p-value indicates the mRNA expression difference between scramble-control and knockdown of each gene. Then, the cell growth was quantified by crystal violet-staining, represented by the images in the middle, with the color intensity quantification indicated by the bar graphs below the images and the difference between scramble-control (S) and knockdown (KD) specified by the p-values. **B**) As a control, siRNA knockdown of two amplified (in

CRCs) DGCs, *LPCAT4* and *ATP6V1B2*, inhibits cell growth. These deleted and amplified genes were also knocked down in HeLa cells and similar results were obtained.

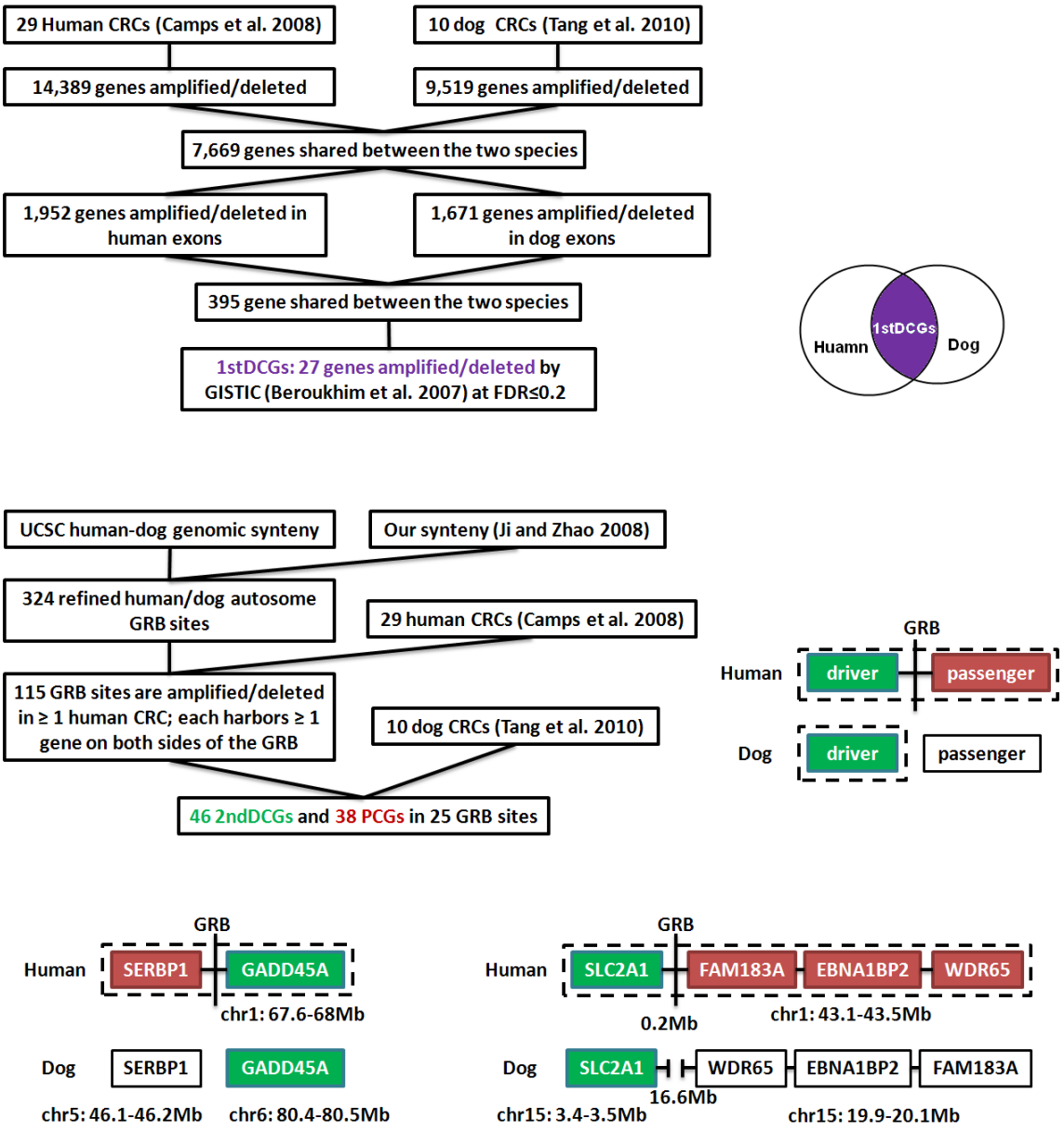
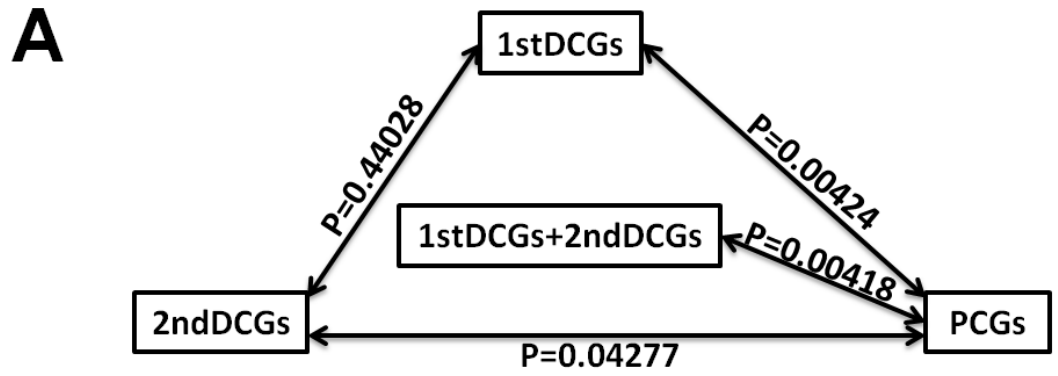
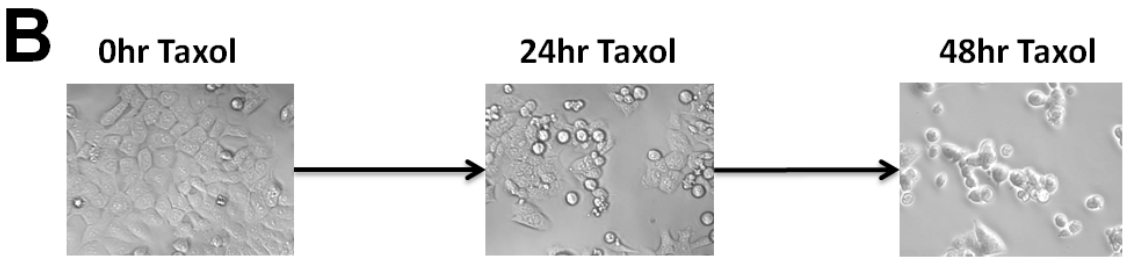

29 Human CRCs (Camps et al. 2008)
10 dog CRCs (Tang et al. 2010)
14,389 genes amplified/deleted
9,519 genes amplified/deleted
7,669 genes shared between the two species
1,952 genes amplified/deleted in human exons
1,671 genes amplified/deleted in dog exons
395 gene shared between the two species
1stDCGs: 27 genes amplified/deleted by GISTIC (Beroukhim et al. 2007) at FDR≤0.2
Huamn
1stDCGs
Dog
UCSC human-dog genomic synteny
Our synteny (Ji and Zhao 2008)
324 refined human/dog autosome GRB sites
29 human CRCs (Camps et al. 2008)
115 GRB sites are amplified/deleted in ≥ 1 human CRC; each harbors ≥ 1 gene on both sides of the GRB
10 dog CRCs (Tang et al. 2010)
46 2ndDCGs and 38 PCGs in 25 GRB sites
GRB
Human
driver
passenger
Dog
driver
passenger
GRB
Human
SERBP1
GADD45A
chr1: 67.6-68Mb
Dog
SERBP1
GADD45A
chr5: 46.1-46.2Mb
chr6: 80.4-80.5Mb
GRB
Human
SLC2A1
FAM183A
EBNA1BP2
WDR65
0.2Mb
chr1: 43.1-43.5Mb
Dog
SLC2A1
WDR65
EBNA1BP2
FAM183A
16.6Mb
chr15: 3.4-3.5Mb
chr15: 19.9-20.1Mb

Figure 3.1





p values of expression change

	0hr vs 24hr Taxol	24h vs 48hr Taxol
1stDCGs	0.009 ↓	0.053 ↑
2ndDCGs	0.014 ↓	0.590
PCGs	0.183	0.452

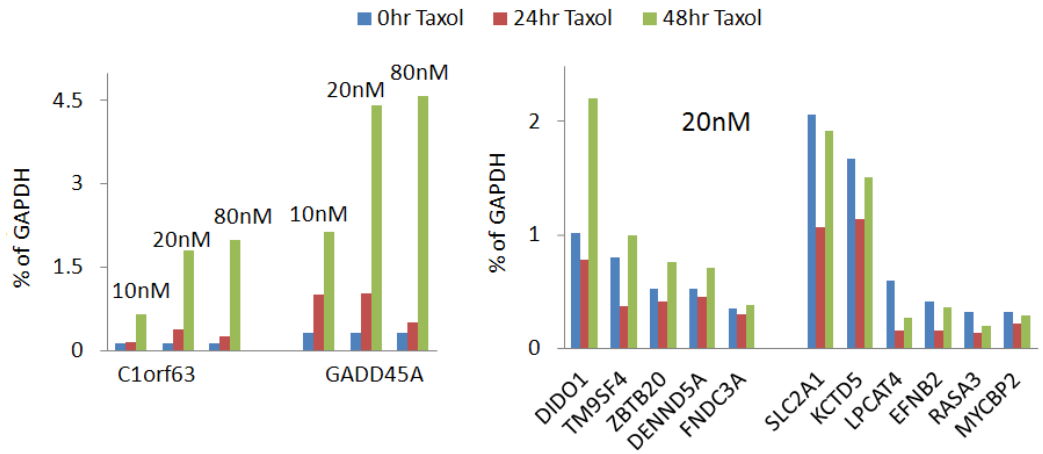
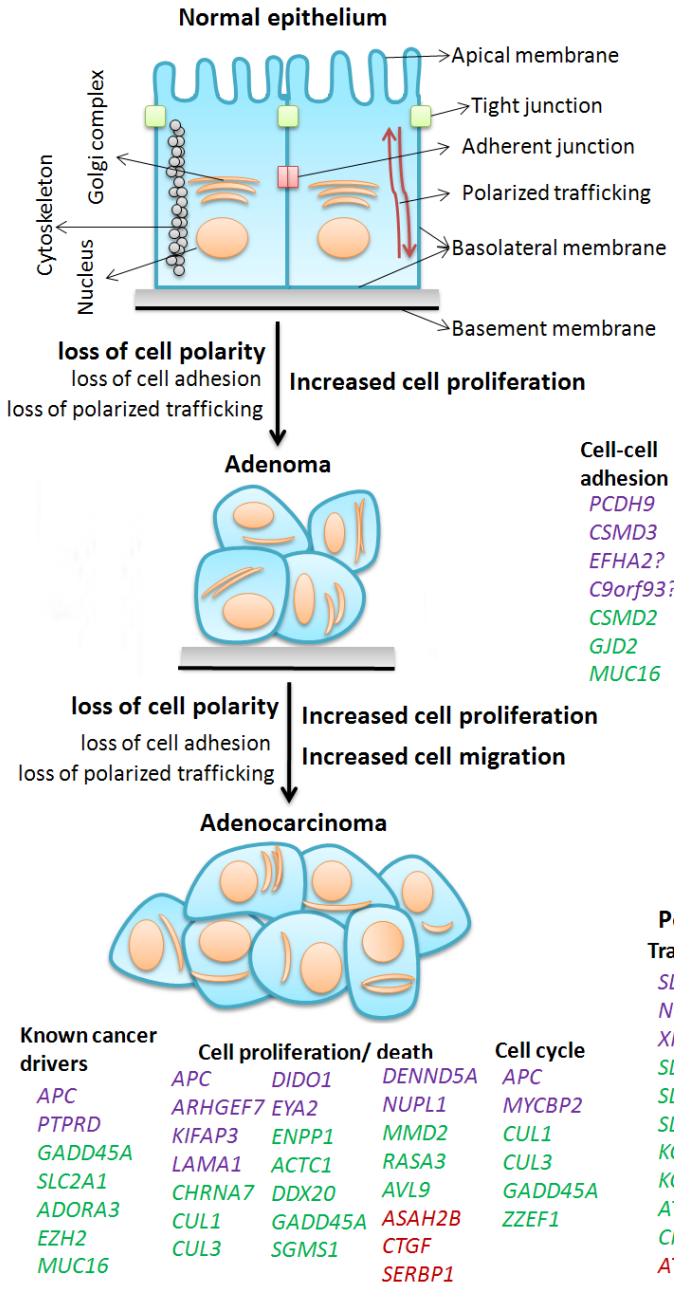


Figure 3.2



Known polarity genes

APC , MPP7, ARHGEF7, LAMA1, PCDH9, FREM2, KIFAP3, DLGAP2, RAP1A, SLC12A6, SLC2A1, ACTC1, ATP6V1B2, ENPP1

MET/EMT genes

EFNB2 , EYA2

Genes located in polarized membrane

Apical:	*ATP6V1B2, MUC16*
Tight junction:	*APC, MPP7*
Adherent juction:	*ARHGEF7*
Basolateral:	*ENPP1 , SLC12A6, SLC2A1*
Basement:	*FREM2, LAMA1*

Adhesion/motility

Cell-cell adhesion	Cell-substratum adhesion	Signaling for adhesion	Cytoskeleton
PCDH9	*LAMA1*	*APC*	*PHACTR3*
CSMD3	*FREM2*	*MPP7*	*MYCBP2*
EFHA2?	*FNDC3A*	*ARHGEF7*	*KIFAP3*
C9orf93?	*PTPRD?*	*DLGAP2*	*SYNPO2L*
CSMD2	*EPDR1*	*RAP1A*	*MYOZ1*
GJD2	*TMPRSS11B*	*SFRP4*	*ACTC1*
MUC16	*TMPRSS11F*	*ARHGAP11A*	*GADD45A*
	AKR1B1?		*DDX20*
	AKR1B10?		*ZZEF1?*
	AKR1B15?		*ACTL9?*
	TXNDC3?		
	CTGF		
	CHIA		
	PRSS33?		

Polarized trafficking/transport

Transport	Trafficking	Signaling for trafficking
SLC7A1	*NPC1*	*DENND5A*
NUPL1	*TM9SF4?*	*ENPP1*
XKR9?	*AVL9*	*ADORA3*
SLC2A1	*MLANA*	*AGAP5*
SLC12A6	*ANKFY1?*	*RASA3*
SLC18A1	*SGMS1?*	*SCG5*
KCND3	*GOLGA8A*	
KCTD5	*GOLGA8B*	
ATP6V1B2	*SEC24C*	
CHRNA7		
ATP2A3?		

Known cancer drivers

APC
PTPRD
GADD45A
SLC2A1
ADORA3
EZH2
MUC16

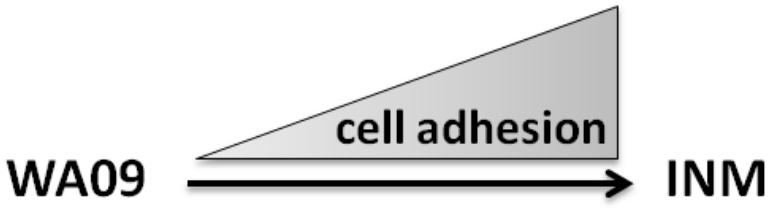
Cell proliferation/ death

APC
ARHGEF7
KIFAP3
LAMA1
CHRNA7
CUL1
CUL3
DIDO1
EYA2
ENPP1
ACTC1
DDX20
GADD45A
SGMS1
DENND5A
NUPL1
MMD2
RASA3
AVL9
ASAH2B
CTGF
SERBP1

Cell cycle

APC
MYCBP2
CUL1
CUL3
GADD45A
ZZEF1

Figure 3.3



P values of expression change

	1stDCGs	2ndDCGs	PCGs
Amplification	0.16235	0.29855	0.24105
Deletion	0.02005 ↑	0.05615 ↑	0.44245

Figure 3.4

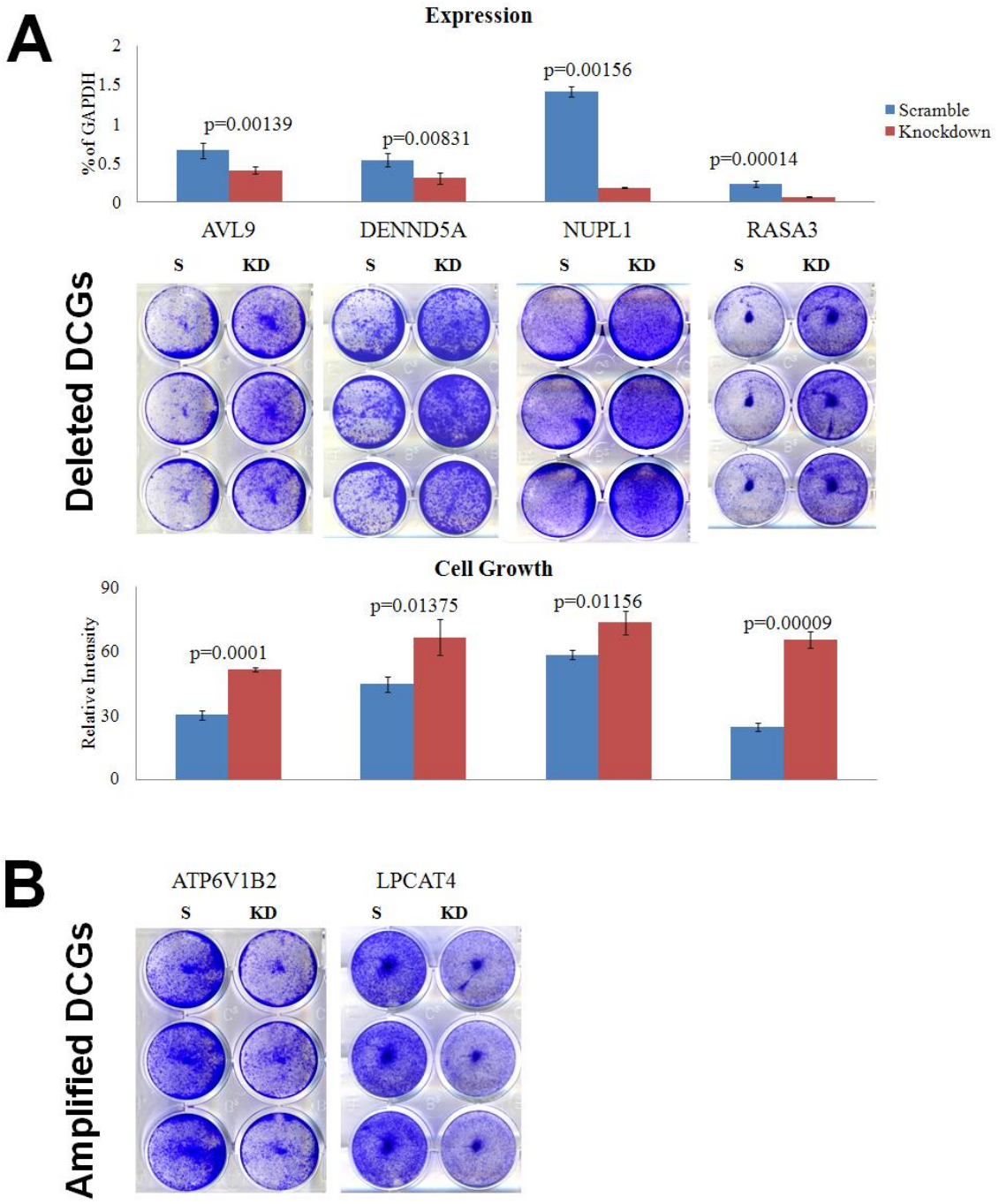

A
Expression
% of GAPDH
2
1.5
1
0.5
0
p=0.00139
p=0.00831
p=0.00156
p=0.00014
Scramble
Knockdown
AVL9
DENND5A
NUPL1
RASA3
S
KD
Deleted DCGs
Cell Growth
Relative Intensity
90
60
30
0
p=0.0001
p=0.01375
p=0.01156
p=0.00009
B
Amplified DCGs
ATP6V1B2
LPCAT4

Figure 3.5

REFERENCES

Amann T, Hellerbrand C. 2009. GLUT1 as a therapeutic target in hepatocellular carcinoma. *Expert opinion on therapeutic targets* **13**(12): 1411-1427.

Argueso P, Guzman-Aranguez A, Mantelli F, Cao Z, Ricciuto J, Panjwani N. 2009. Association of cell surface mucins with galectin-3 contributes to the ocular surface epithelial barrier. *The Journal of biological chemistry* **284**(34): 23037-23045.

Bachmann O, Juric M, Seidler U, Manns MP, Yu H. 2011. Basolateral ion transporters involved in colonic epithelial electrolyte absorption, anion secretion and cellular homeostasis. *Acta Physiol (Oxf)* **201**(1): 33-46.

Barrios A, Poole RJ, Durbin L, Brennan C, Holder N, Wilson SW. 2003. Eph/Ephrin signaling regulates the mesenchymal-to-epithelial transition of the paraxial mesoderm during somite morphogenesis. *Current biology : CB* **13**(18): 1571-1582.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**(7): 1159-1170.

Belleannee C, Da Silva N, Shum WW, Brown D, Breton S. 2010. Role of purinergic signaling pathways in V-ATPase recruitment to apical membrane of acidifying epididymal clear cells. *American journal of physiology Cell physiology* **298**(4): C817-830.

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S et al. 2007. Assessing the significance of

chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**(50): 20007-20012.

Bouanene H, Miled A. 2010. Conflicting views on the molecular structure of the cancer antigen CA125/MUC16. *Disease markers* **28**(6): 385-394.

Bryant DM, Mostov KE. 2008. From cells to organs: building polarized tissue. *Nature reviews Molecular cell biology* **9**(11): 887-901.

Camps J, Grade M, Nguyen QT, Hormann P, Becker S, Hummon AB, Rodriguez V, Chandrasekharappa S, Chen Y, Difilippantonio MJ et al. 2008. Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. *Cancer research* **68**(5): 1284-1295.

Cyranoski D. 2010. Genetics: Pet project. *Nature* **466**(7310): 1036-1038.

Farabaugh SM, Micalizzi DS, Jedlicka P, Zhao R, Ford HL. 2011. Eya2 is required to mediate the pro-metastatic functions of Six1 via the induction of TGF-beta signaling, epithelial-mesenchymal transition, and cancer stem cell properties. *Oncogene*.

Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. 1993. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**(5): 1027-1038.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**(Database issue): D945-950.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nature reviews Cancer* **4**(3): 177-183.

Gibbs RA. 2011. Genome-sequencing anniversary. Bringing genomics and genetics back together. *Science* **331**(6017): 548.

Haber DA, Settleman J. 2007. Cancer: drivers and passengers. *Nature* **446**(7132): 145-146.

Harsay E, Schekman R. 2007. Avl9p, a member of a novel protein superfamily, functions in the late secretory pathway. *Molecular biology of the cell* **18**(4): 1203-1219.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**(1): 44-57.

Hudson TJ Anderson W Artez A Barker AD Bell C Bernabe RR Bhan MK Calvo F Eerola I Gerhard DS et al. 2010. International network of cancer genome projects. *Nature* **464**(7291): 993-998.

Iden S, Collard JG. 2008. Crosstalk between small GTPases and polarity proteins in cell polarization. *Nature reviews Molecular cell biology* **9**(11): 846-859.

Jacobson KA, Klutz AM, Tosh DK, Ivanov AA, Preti D, Baraldi PG. 2009. Medicinal chemistry of the A3 adenosine receptor: agonists, antagonists, and receptor engineering. *Handbook of experimental pharmacology*(193): 123-159.

Ji X, Tang J, Halberg R, Busam D, Ferriera S, Pena MM, Venkataramu C, Yeatman TJ, Zhao S. 2010. Distinguishing between cancer driver and passenger gene alteration candidates via cross-species comparison: a pilot study. *BMC cancer* **10**: 426.

Ji X, Zhao S. 2008. DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. *Genomics* **91**(3): 249-258.

Jordan MA, Wilson L. 2004. Microtubules as a target for anticancer drugs. *Nature reviews Cancer* **4**(4): 253-265.

Kalluri R, Weinberg RA. 2009. The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation* **119**(6): 1420-1428.

Kinzler KW, Vogelstein B. 1996. Lessons from hereditary colorectal cancer. *Cell* **87**(2): 159-170.

Krais AM, Hautefeuille AH, Cros MP, Krutovskikh V, Tournier JM, Birembaut P, Thepot A, Paliwal A, Herceg Z, Boffetta P et al. 2011. CHRNA5 as negative regulator of nicotine signaling in normal and cancer bronchial cells: effects on motility, migration and p63 expression. *Carcinogenesis* **32**(9): 1388-1395.

Ledford H. 2010. Big science: The cancer genome challenge. *Nature* **464**(7291): 972-974.

Lee M, Vasioukhin V. 2008. Cell polarity and cancer--cell and tissue polarity as a non-canonical tumor suppressor. *Journal of cell science* **121**(Pt 8): 1141-1150.

Levy F, Muehlethaler K, Salvi S, Peitrequin AL, Lindholm CK, Cerottini JC, Rimoldi D. 2005. Ubiquitylation of a melanosomal protein by HECT-E3 ligases serves as sorting signal for lysosomal degradation. *Molecular biology of the cell* **16**(4): 1777-1787.

Li R, Gundersen GG. 2008. Beyond polymer polarity: how the cytoskeleton builds a polarized cell. *Nature reviews Molecular cell biology* **9**(11): 860-873.

Lin J, Gan CM, Zhang X, Jones S, Sjoblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B et al. 2007. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome research* **17**(9): 1304-1318.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(7069): 803-819.

Long BH, Fairchild CR. 1994. Paclitaxel inhibits progression of mitotic cells to G1 phase by interference with spindle formation without affecting other microtubule functions during anaphase and telephase. *Cancer research* **54**(16): 4355-4361.

Lott ST, Chen N, Chandler DS, Yang Q, Wang L, Rodriguez M, Xie H, Balasenthil S, Buchholz TA, Sahin AA et al. 2009. DEAR1 is a dominant regulator of acinar morphogenesis and an independent predictor of local recurrence-free survival in early-onset breast cancer. *PLoS medicine* **6**(5): e1000068.

Nelson WJ. 2003. Adaptation of core mechanisms to generate cell polarity. *Nature* **422**(6933): 766-774.

Pasche B, Myers RM. 2009. One step forward toward identification of the genetic signature of glioblastomas. *JAMA : the journal of the American Medical Association* **302**(3): 325-326.

Pontius JU, Mullikin JC, Smith DR, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B, Volfovsky N et al. 2007. Initial sequence and comparative analysis of the cat genome. *Genome research* **17**(11): 1675-1689.

Reinhardt HC, Hasskamp P, Schmedding I, Morandell S, van Vugt MA, Wang X, Linding R, Ong SE, Weaver D, Carr SA et al. 2010. DNA damage activates a

spatially distinct late cytoplasmic cell-cycle checkpoint network controlled by MK2-mediated RNA stabilization. *Molecular cell* **40**(1): 34-49.

Rosemary Siafakas A, Richardson DR. 2009. Growth arrest and DNA damage-45 alpha (GADD45alpha). *The international journal of biochemistry & cell biology* **41**(5): 986-989.

Rowell JL, McCarthy DO, Alvarez CE. 2011. Dog models of naturally occurring cancer. *Trends in molecular medicine* **17**(7): 380-388.

Royer C, Lu X. 2011. Epithelial cell polarity: a major gatekeeper against cancer? *Cell death and differentiation* **18**(9): 1470-1477.

Sawyers CL. 1999. Chronic myeloid leukemia. *The New England journal of medicine* **340**(17): 1330-1340.

Sheikh MS, Hollander MC, Fornance AJ, Jr. 2000. Role of Gadd45 in apoptosis. *Biochemical pharmacology* **59**(1): 43-45.

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**(5797): 268-274.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**(6942): 825-837.

Sottocornola R, Royer C, Vives V, Tordella L, Zhong S, Wang Y, Ratnayaka I, Shipman M, Cheung A, Gaston-Massuet C et al. 2010. ASPP2 binds Par-3 and controls the

polarity and proliferation of neural progenitors during CNS development. *Developmental cell* **19**(1): 126-137.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**(7239): 719-724.

Tang J, Le S, Sun L, Yan X, Zhang M, Macleod J, Leroy B, Northrup N, Ellis A, Yeatman TJ et al. 2010. Copy number abnormalities in sporadic canine colorectal cancers. *Genome research* **20**(3): 341-350.

Veeriah S, Brennan C, Meng S, Singh B, Fagin JA, Solit DB, Paty PB, Rohle D, Vivanco I, Chmielecki J et al. 2009. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proceedings of the National Academy of Sciences of the United States of America* **106**(23): 9435-9440.

Yoshimura S, Gerondopoulos A, Linford A, Rigden DJ, Barr FA. 2010. Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors. *The Journal of cell biology* **191**(2): 367-381.

Youmans L, Taylor C, Shin E, Harrell A, Ellis AE, Seguin B, Ji X, Zhao S. Frequent alteration of the tumor suppressor gene APC in sporadic canine colorectal cancers. *PloS ONE, in revision*.

Yvon AM, Wadsworth P, Jordan MA. 1999. Taxol suppresses dynamics of individual microtubules in living human tumor cells. *Molecular biology of the cell* **10**(4): 947-959.

CHAPTER 4

CONCLUSIONS

Consistent with the reported anatomic and clinical similarities, our initial characterization of genomic instability in ten spontaneously occurring dog tumors revealed for the first time the genetic and molecular similarities between sporadic human and dog CRCs. All ten canine tumors investigated exhibited CIN, with the extent of CIN correlating with the tumor's progression stage, origin, and likely MSI status reminiscent of human CRCs. However, the MSI phenotype was found in only one canine tumor. This is consistent with human sporadic CRCs, where the majority (87%) display CIN and the minority (13%) display MSI. In addition, our analyses indicate the corresponding human CRC pathways may also be altered in the dog cancers and that tumors from the two species were clustered according to the tumor subtypes, but not according to the species. These observations strongly suggest that human and dog CRCs share similar molecular and genetic pathways of cancer development and progression, although more dog tumors merit investigation.

We discovered that the species-specific CNAs localize to genomic regions that are evolutionarily more unstable compared with the CNAs shared between the two species. Our finding raises the possibility that the human-specific CNAs may be a consequence of cancer development and progression, rather than a cause. This is because

these genomic regions are intrinsically more unstable and are thus more prone to changes when the genome becomes increasingly more unstable as cancer progresses, compared with other genomic sites. Thus, these human-specific CNAs may be less significant than the shared CNAs in CRC etiology, especially if both species are clearly shown to follow similar molecular and genetic pathways of cancer development and progression.

As an expansion, we used sporadic dog CRCs as a model to identify bona fide CRC gene candidates and successfully demonstrated that our human-dog comparison strategy for CRC driver-passenger distinction is valid. The 73 identified DCGs significantly differ from the identified 38 PCGs in every analysis performed. More importantly, we discovered that, while PCGs are not enriched in any specific functions, DCGs hold significantly enhanced functionalities that are closely associated with epithelial cell polarity establishment and maintenance. Meanwhile, many DCGs participate in processes that regulate cell proliferation and death, and/or the cell cycle. Hence, this pilot study indicates that, in sporadic CRCs of both species, genes that establish and maintain epithelial cell polarity not only play a critical role in preventing cancer cell invasion and spreading, but also likely serve as tumor suppressors by modulating cell growth.

Unlike previously published papers that are limited to the mere identification of genes recurrently altered in CRCs [e.g., (Sjoblom et al. 2006)], this human-dog comparison approach allowed us to further classify these genes as either cancer driver candidates or passenger candidates, effectively narrowing down the targets for

downstream functional validation. Expanding this pilot study to a larger scale would likely make many more driver-passenger distinctions and address key questions regarding cancer pathogenesis and epithelial cell polarity control.

REFERENCES

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N et al. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314(5797): 268-274.