

Copy number abnormalities in sporadic canine colorectal cancers

Jie Tang,¹ Shoshona Le,² Liang Sun,³ Xiuzhen Yan,¹ Mucheng Zhang,¹ Jennifer MacLeod,⁴ Bruce LeRoy,⁵ Nicole Northrup,⁵ Angela Ellis,⁵ Timothy J. Yeatman,⁶ Yanchun Liang,³ Michael E. Zwick,² and Shaying Zhao^{1,7}

¹Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602, USA;

²Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30602, USA; ³College of Computer Science and Technology, Jilin University, Changchun 130021, China; ⁴School of Veterinary Medicine, University of California Davis, California 95616, USA; ⁵College of Veterinary Medicine, University of Georgia, Athens, Georgia 30602, USA; ⁶Departments of Surgery, Pathology, and Biostatistics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida 33612, USA

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. Last, 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.

[Supplemental material is available online at <http://www.genome.org>. The aCGH data have been submitted to the NCBI Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) under accession no. GSE19318.]

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 (<http://cancergenome.nih.gov/>), the International Cancer Genome Consortium (<http://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., Beroukhim et al. 2007; Greenman et al. 2007; Jones et al. 2008).

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) (Fig. 1). In our study, evolutionarily unstable sites are defined as regions enriched with interspecies genomic rearrangement breakpoints (ISGRBPs) (Pevzner and Tesler 2003; Zhao et al. 2004) and segmental duplications (SDs) (Bailey et al. 2002, 2004; Turner et al. 2007; Kidd et al. 2008).

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

⁷Corresponding author.

E-mail szhao@bmb.uga.edu; fax (706) 542-1738.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.092726.109>.

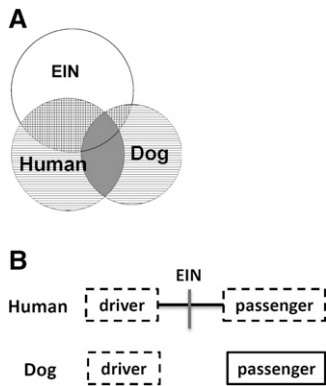


Figure 1. (A) Cross-species comparison for causative (or driver) aberration identification. Once we demonstrate that the same types of cancer from the human and the dog share similar molecular and genetic pathways of cancer development and progression, we will consider abnormalities recurrent between the two species as driver candidates (solid gray area), and those found in only one species and falling in evolutionarily unstable sites (EIN sites) as passenger candidates (small squared areas). Those in the parallel-lined areas need further studies. (B) 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. This shows that two genes, which are nearby in the human genome but distant in the dog genome, are both disrupted in the human cancer (boxed with broken lines). However, in the dog cancer, only one gene is disrupted, which will be considered as driver, and the other is intact (boxed with unbroken lines), which will be deemed as passenger.

are recurrent not only among different cases within the same species, but also among different species. Compared with single-species approaches, this multispecies 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 (Fig. 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 and 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-beta, p53, and others that play fundamental roles in CRC development and progression (Kinzler and Vogelstein 1996; Gallahan and Callahan 1997; Rajagopalan et al. 2003; Grady 2004; Sancho et al. 2004; Kaiser et al. 2007).

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 (Fig. 2), as well as tumors of epithelial origin (adenomas and adenocarcinomas) and nonepithelial origin (a mast cell tumor; a leiomyosarcoma that arose from muscle cells). As summarized in Table 1, we identified between 2000 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 Fig. 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*, *TGFBR2*, *TGFBR1*, *PTEN*, and *TP53*), as well as two deleted regions on chromosomes 5 and 9. Through *t*-tests (see Methods), we confirmed deletions in *APC*, *PTEN*, *SMAD3*, *SMAD2*, *SMAD4*, *TP53*, and the regions on chromosomes 5 and 9, as well as amplifications in *TGFBR2* and *GSK3B* (*P*-values

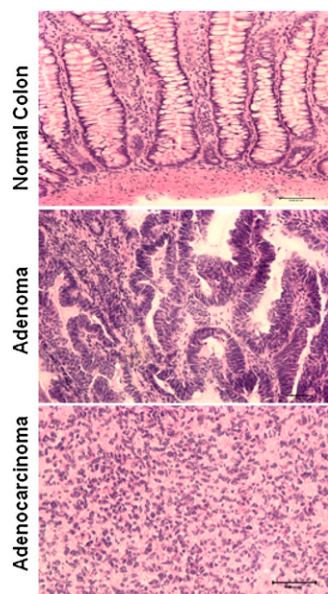


Figure 2. Cryosectioning and H&E staining of dog colon tumor and normal tissue samples. The images represent a normal colon tissue (top), an adenoma (middle), and an adenocarcinoma (bottom).

ranged from 0.001 to <0.1), indicating the accuracy of our aCGH analyses.

The amount of CNAs directly correlated with the tumor's progression stage, origin, and MSI status

As shown in Table 1, the total amount of CNAs identified varied significantly among the 10 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 likely the MSI status of the tumors. First, tumors with fewer CNAs are at early stages (~146–285 Mb for adenomas), whereas tumors with a greater number of CNAs are at late stages (~350–550 Mb for adenocarcinomas) (Table 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 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 (146 Mb vs. 285 and 253 Mb).

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 ($i = 1, 2, 3, \dots, 10$), respectively, with their values indicated in Supplemental Table s1. We found that the tumor T3 (Table 1) is an outlier in this analysis; if excluding T3, the observed data in Supplemental Table s1 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 ($P = 0.0001$). We understand that we had a small sample size here (only nine tumors). According to Knofczynski (2008), the sample size requirement of multiple linear regressions varies with R^2 and, for $R^2 > 0.9$, a good prediction with three variables requires at least nine samples. Hence, our analysis satisfied this minimal sample size requirement.

Tumor clustering

We also performed clustering analyses using pairwise overlapping CNAs. Both the minimum spanning tree (MST) (Kruskal 1956; Prim 1957; Cormen et al. 2001) and top-down clustering strategies (Liu et al. 2006) (see Methods), as well as hierarchical clustering (Hastie et al. 2009), yielded the same structures as shown in Figure 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 nonepithelial 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 (Supplemental Fig. s1), consistent with the multiple linear regression results described above. However, within the same subtype of adenocarcinomas, tumors

Table 1. CNAs identified in 10 dog colorectal tumors

Dog tumors ^a	Total CNA no.	Average probe no.	Average log ₂ ratio	Average CNA size (kb)	Total CNA size on dog genome (Mb)	Total CNA size on human genome (Mb)	Total no. of genes
T1, large adenoma, MSI-low	2391	11	0.188	61	146	151	2136
T4, adenoma	5708	9	0.201	50	285	315	3783
T6, adenoma	5880	8	0.205	43	253	277	3555
T3, adenocarcinoma	7022	9	0.211	50	349	382	4305
T8, adenocarcinoma	9998	10	0.234	54	545	599	6577
T9, adenocarcinoma	9389	11	0.236	58	550	601	6739
T10, adenocarcinoma	9588	10	0.238	54	514	559	6347
T11, adenocarcinoma	9107	11	0.228	60	551	609	5715
T5, metastatic mast cell tumor	4974	8	0.193	40	197	218	3066
T7, Leiomyosarcoma	3486	8	0.191	39	138	152	2198

Chromosomes X, Y, M, and UN were not included.

^aAdenomas (T1, T4, and T6) are early-stage tumors, with T1 being more advanced than T4 and T6 and exhibiting a weak MSI phenotype. Adenocarcinomas (T3 and 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.

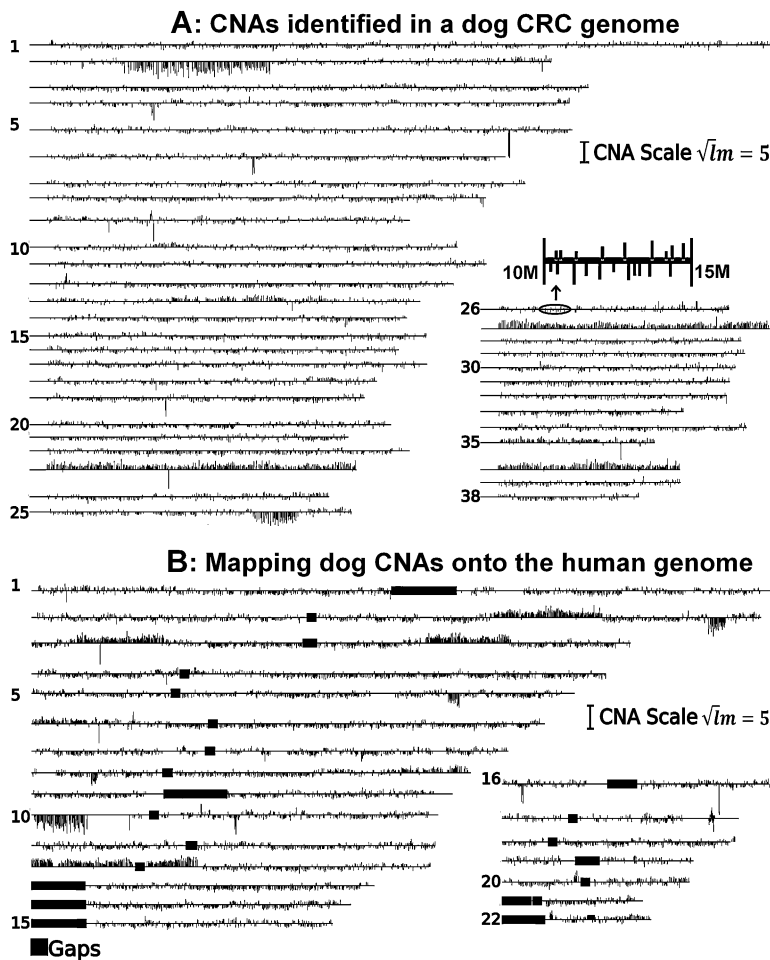


Figure 3. (A) CNAs identified in the genome of a dog adenocarcinoma (T11). The identified CNAs (3396 gains and 5711 losses) amount to 551 Mb (22% of the dog genome) (Table 1). Each line represents a dog chromosome with its chromosome number indicated on the *right*. Vertical lines shown *above/below* the chromosomes represent gains/losses, respectively, with their length calculated based on \sqrt{lm} , 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 9107 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.

clustered based on not only the total CNA size, but also the CNA pattern (Supplemental Fig. s1).

Genes and pathways disrupted in the dog colorectal tumors

Mapping of dog CNAs onto the human genome

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 1; for one example, see Fig. 3). 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 dog–human synteny is not yet resolved (12%–26%; see Supplemental Table s2 for more information). 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.

Known human CRC pathways were likely to be disrupted recurrently in dog tumors

Using the mapping information described above, we found about 2000–4000 genes for early-stage/nonepithelial/probable MSI-low tumors and 5000–7500 genes for the late stage epithelial tumors (or adenocarcinomas) (Table 1). For 73%–86% of these genes, we were able to annotate their functions and classify them using the Gene Ontology (GO) 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 (see Supplemental Fig s2).

Importantly, we asked if any of these genes participate in pathways known to be involved in human CRC development and progression (Kinzler and Vogelstein 1996; Gallahan and Callahan 1997; Rajagopalan et al. 2003; Grady 2004; Sancho et al. 2004; Kaiser et al. 2007). These included well-characterized signaling pathways of Wnt, TGF- β , TP53, and MAPK, as well as pathways controlling cell growth and apoptosis, and cell cycle (see <http://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 2A). Table 2B lists those genes that were recurrently disrupted among the dog tumors, identified following a strategy described by Beroukhi et al. (2007) (see 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

Human and dog tumors of similar subtypes were clustered together

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

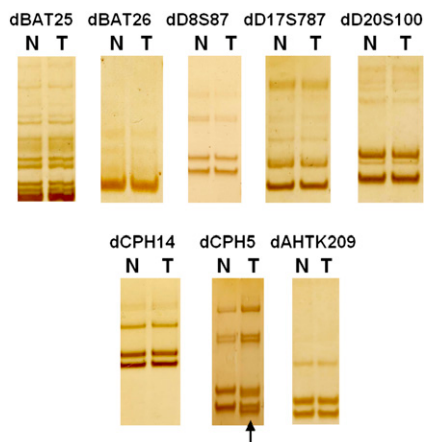


Figure 4. MSI assay of the dog tumors. A total of eight dog microsatellite loci were analyzed as described in the text, with the *top* five being homologs of frequently used human MSI markers (Boland et al. 1998) and the *bottom* three being dinucleotide markers used by McNeil et al. (2007) to determine MSI status in canine mammary gland neoplasia. The extra band that the tumor displayed for the locus CPAS was indicated by an arrow. T, tumor; N, its matching normal.

(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 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., Douglas et al. 2004; Nakao et al. 2004; Camps et al. 2006). Like the dog analyses described before, these results also demonstrate the accuracy of our aCGH analyses.

Importantly, we found that ~12%–33% of CNAs of the human MSI-none tumor colocalized 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 ($P < 10^{-4}$; see Supplemental Table s3). This close relationship was also demonstrated by the clustering analyses. As shown in Figure 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 (Fig. 5B). These results provide another piece of evidence demonstrating the molecular and genetic similarity between human and canine CRCs.

Human-specific CNAs locate in evolutionarily more unstable regions compared with the shared CNAs

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 ISGRBPs (mouse, rat, and dog genomic rearrangement breakpoints on the human genome) and SDs in them. We found that the shared CNAs contain two to three times fewer ISGRBPs and up to three times fewer SDs compared with the human-specific CNVs (Supplemental Table s4). 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 the same efficiency as the total CNAs (Supplemental Table s5), 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 ($\geq 20\%$ 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 $\geq 20\%$ of their genomic sequences were amplified or deleted in at least two dog tumors ($\geq 20\%$ of the total sample size) to compare with the selected human CNAs.

As summarized in Table 3, we found that ~64% of the human CNAs and 56% of the dog CNAs colocalize 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

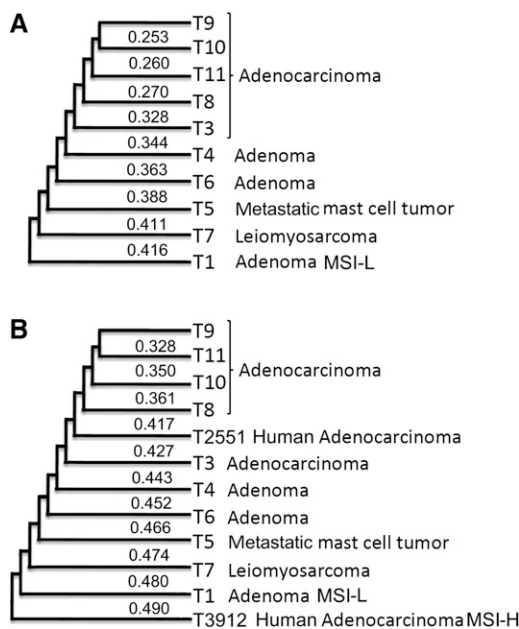


Figure 5. (A) Clustering of the dog tumors. The tree was constructed by MST or top-down clustering as described in the text (both strategies yielded the same tree), with the sample information for each tumor shown on the right. The number for each branch represents the distance $d(X, Y)$ between the two clusters X and Y involved, calculated by $d(X, Y) = \sum_{T_i \in X, T_j \in Y} d_{ij} / (|X||Y|)$, 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. (B) 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 Fig. 3) the human genome. MSI-L: MSI-low; MSI-H: MSI-high.

Table 2A. Observed and predicted numbers of disrupted genes that participate in known human CRC pathways

Tumor ^a	Observed no.	Predicted no. ^b
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

The one-tail *t*-test indicates that observed gene numbers are significantly more than predicted values ($P = 0.0004$).

^aT1–T11 are dog tumors (see Table 1 for more detailed description), whereas T2551 is a human MSI-none adenocarcinoma, and T3912 is a human MSI-high adenocarcinoma.

^bThe 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.

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., Moser et al. 1993; Vail and MacEwen 2000; Rosol et al. 2003; Hansen and Khanna 2004; Hinoi et al. 2007; Maser et al. 2007). 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 (Vail and MacEwen 2000; Michell 2004; Argyle 2005, 2009; Khanna et al. 2006; Paoloni and Khanna 2007, 2008; LeRoy and 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 (Fig. 1).

Consistent with the reported anatomic and clinical similarities, our initial characterization of genomic instability in 10 spontaneously occurring dog tumors revealed for the first time the genetic and molecular similarities between sporadic

human and dog CRCs. All 10 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 (Fig. 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 (McNiell et al. 2007), have not been extensively tested. Importantly, we do not know whether disruption of the dog locus shown in Figure 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 (6.2–12.7 Mb) 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 10 dog tumors investigated (Table 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 nine 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

Table 2B. Recurrently disrupted dog homologs of genes participating in known human CRC pathways

Pathways	Genes ^a
Apoptosis	<i>BCL2, FADD, PIK3R3</i>
Cell cycle	<i>ABL1, ANAPC10, MAD1L1, MCMS, PRKDC</i>
MAPK signaling	<i>CACNA1C, CACNA1D, CACNA1I, CACNA2D3, CACNA2D4, CACNB2, FGF10, FGF12, FGF14, MAP2K5, MAP3K7IP1, MAP3K7IP2, MECOM, NTRK2, PDGFRA, PPM1B, RAPGEF2, RASGRF1, RPS6KA2, RPS6KA5, STK3, TAOK1, ZAK</i>
P53 signaling	<i>APAF1, CCND1, MDM4, PPM1D, PTEN, TP73</i>
TGF-beta signaling	<i>ACVR1, ACVR2A, BMP6, BMPR1B, GDF6, LTBP1</i>
Wnt signaling	<i>CTBP2, DKK2, MAPK10, NKD1, NLK, PLCB1, PLCB2, PLCB4, PPP2R5E</i>

^aShown here are significantly disrupted genes 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) (see Supplemental materials).

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

effort reported that, in the genome of the breast cancer cell line MCF7, dispersed rearrangement breakpoint regions are significantly enriched with SDs (5.2×), whereas clustered breakpoint 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).

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 University. Tissue samples were obtained during surgery. After washing in phosphate-buffered saline, they were snap-frozen in liquid nitrogen for 10 min and then stored at -80°C 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- μm slices, which were stained with hematoxylin and eosin Y (Fig. 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 five 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 three additional dinucleotide markers (CPH14, CPH5, and AHTK209), based on a previous study of canine mammary gland neoplasia (McNiel et al. 2007). The primers used to amplify these markers were listed in Supplemental Table s6.

PCR reactions were carried out in a 10- μ L reaction volume containing 25 ng of genomic DNA, 25 μ M primers, and 5 μ L of 2 \times iQ supermix (Bio-Rad). The reactions were performed under the following conditions: 95°C for 5 min, 30 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 30 sec, and a final extension at 72°C 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 the 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 Arrays User's Guide." A detailed description of the protocol is provided in the Supplemental materials. 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 one 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 one 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 five 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 and 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 <http://www.bmb.uga.edu/szhaoh>.

Multiple linear regression analysis was performed using the statistic software JMP8 (<http://www.jmp.com>; SAS Institute). 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 nonepithelial origin and 1 for epithelial origin), and the MSI status (0 for MSI+ and 1 for MSI-) of each tumor (see Supplemental Table s1).

qPCR analysis

qPCR reactions were performed in triplicate with each well containing 10 μ L of iQ SYBER Green Supermix from Bio-Rad, 500 nM primers each, and 10 ng of genomic DNA in a total reaction volume of 20 μ L, with an iCycler iQ Real-Time PCR machine. The amplification condition was: 95°C for 10 sec, 65°C for 45 sec, and 78°C for 20 sec, 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 Ct value means fewer templates) was collected for each reaction. 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 10 tumors studied, C_i and C_j represented the total genomic size of CNAs in T_i and 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 10 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 (Kruskal 1956; Prim 1957; Cormen et al. 2001). Last, 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

$$IC = \frac{\sum_{r=1}^2 \sum_{i < j, T_i, T_j \in C_r} S_{ij}}{b_r} \text{ and}$$

$$IS = \frac{\sum_{T_i \in C_1, T_j \in C_2} S_{ij}}{b_1 b_2},$$

where C_1, C_2 represents the two clusters that, respectively, harbor a total number of b_1, b_2 tumors. Since both measures were computed with the pairwise similarity among the tumors, higher values of IC and lower values of IS would represent better clustering quality. Consequently, in our current implementation, if 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 (<http://www.genome.ucsc.edu>). Through database cross-linking, functions of the identified genes were annotated based on the GOA database, version 1.109 (<http://www.ebi.ac.uk/GOA/>). The subsequent GO slim and classification analysis at the biological process level was achieved by Blast2Go, version 2.35 (<http://www.blast2go.de/>). Information on known human CRC-related pathways was obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) release 50.0 (<http://www.genome.jp/kegg>). Genes recurrently disrupted in dog tumors (Table 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 Beroukhi et al. (2007) (see Supplemental material). Previously published CNAs of human CRCs from BAC aCGH studies were downloaded from the NCBI's SKY/M-FISH and CGH database at <http://www.ncbi.nlm.nih.gov/sky>. Human and dog SD data, as well as human cytogenetic band and genomic sequence coordinate conversion data, were obtained from the annotation data set 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).

Acknowledgments

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

References

Aebbersold R, Auffray C, Baney E, Barillot E, Brazza A, Brett C, Brunak S, Butte A, Califano A, Celis J, et al. 2009. Report on EU-USA workshop:

- How systems biology can advance cancer research (27 October 2008). *Mol Oncol* **3**: 9–17.
- Albertson DG, Collins C, McCormick F, Gray JW. 2003. Chromosome aberrations in solid tumors. *Nat Genet* **34**: 369–376.
- Argyle DJ. 2005. The benefits of comparative medicine—a hundred years to come to our senses. *Vet J* **170**: 147–148.
- Argyle DJ. 2009. Prostate cancer in dogs and men: A unique opportunity to study the disease. *Vet J* **180**: 137–138.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.
- Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol* **5**: R23. <http://genomebiology.com/2004/5/4/R23>.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple-testing. *JR Stat Soc* **57**: 289–300.
- Beroukhi 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. *Proc Natl Acad Sci* **104**: 20007–20012.
- Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, et al. 1998. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* **58**: 5248–5257.
- 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**: 419–428.
- Camps J, Grade M, Nguyen QT, Hörmann 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 Res* **68**: 1284–1295.
- Cormen TH, Leiserson CE, Rivest RL, Stein C. 2001. *Introduction to algorithms*, 2nd ed., The MIT Press, Cambridge and McGraw-Hill, New York.
- Deng L, Zhang Y, Kang J, Liu T, Zhao H, Gao Y, Li C, Pan H, Tang X, Wang D, et al. 2008. An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Hum Mutat* **29**: 1209–1216.
- 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**: 4817–4825.
- Gallahan D, Callahan R. 1997. The mouse mammary tumor associated gene *INT3* is a unique member of the *NOTCH* gene family (*NOTCH4*). *Oncogene* **14**: 1883–1890.
- Grady WM. 2004. Genomic instability and colon cancer. *Cancer Metastasis Rev* **23**: 11–27.
- Greenman C, Stephens P, Smith R, Dalgleish 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**: 153–158.
- Hampton OA, Den Hollander P, Miller CA, Delgado DA, Li J, Coarfa C, Harris RA, Richards S, Scherer SE, Muzny DM, Gibbs RA, 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. **19**: 167–177.
- Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**: 57–70.
- Hansen K, Khanna C. 2004. Spontaneous and genetically engineered animal models: Use in preclinical cancer drug development. *Eur J Cancer* **40**: 858–880.
- Hastie T, Tibshirani R, Friedman J. 2009. Hierarchical clustering. In *The elements of statistical learning*, 2nd ed., pp. 520–528. Springer, New York.
- Haverty PM, Fridlyand J, Li L, Getz G, Beroukhi R, Lohr S, Wu TD, Cavet G, Zhang Z, Chant J. 2008. High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer* **47**: 530–542.
- Hinoi T, Akyol A, Theisen BK, Ferguson DO, Greenson JK, Williams BO, Cho KR, Fearon ER. 2007. Mouse model of colonic adenoma-carcinoma progression based on somatic Apc inactivation. *Cancer Res* **67**: 9721–9730.
- Ji X, Zhao S. 2008. DA and Xiao—two composite and giant LTR-retrotransposon like structures identified on the human genome. *Genomics* **91**: 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**: 1801–1806.

- Kaiser S, Park YK, Franklin JL, Halberg RB, Yu M, Jessen WJ, Freudenberg J, Chen X, Haigis K, Jegga AG, et al. 2007. Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. *Genome Biol* **8**: R131. doi: 10.1186/gb-2007-8-7-r131.
- 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. *Nat Biotechnol* **24**: 1065–1066.
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.
- Kinzler KW, Vogelstein B. 1996. Lessons from hereditary colorectal cancer. *Cell* **87**: 159–170.
- Knofczynski GT. 2008. Sample sizes when using multiple linear regression for prediction. *Educ Psychol Meas* **68**: 431–442.
- Knutsen T, Gobu V, Knaus R, Padilla-Nash H, Augustus M, Strausberg RL, Kirsch IR, Sirotkin K, Ried T. 2005. 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* **44**: 52–64.
- Kruskal JB. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc Am Math Soc* **7**: 48–50.
- Lengauer C, Kinzler KW, Vogelstein B. 1997. Genetic instability in colorectal cancers. *Nature* **386**: 623–627.
- Leroy BE, Northrup N. 2009. Prostate cancer in dogs: Comparative and clinical aspects. *Vet J* **180**: 149–162.
- Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ III, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.
- Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M. 2006. Distance-based clustering of CGH data. *Bioinformatics* **22**: 1971–1978.
- Maser RS, Choudhury B, Campbell PJ, Feng B, Wong KK, Protopopov A, O'Neil J, Gutierrez A, Ivanova E, Perna I, et al. 2007. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* **447**: 966–971.
- McNiel EA, Griffin KL, Mellett AM, Madrill NJ, Mickelson JR. 2007. Microsatellite instability in canine mammary gland tumors. *J Vet Intern Med* **21**: 1034–1040.
- Michell AR. 2004. Comparative clinical science: The medicine of the future. *Vet J* **170**: 153–162.
- Moser AR, Mattes EM, Dove WF, Lindstrom MJ, Haag JD, Gould MN. 1993. *Apc^{Min}*, a mutation in the murine *Apc* gene, predisposes to mammary carcinomas and focal alveolar hyperplasias. *Proc Natl Acad Sci* **90**: 8977–8981.
- Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, Lafuente A, Wiencke JW, Terdiman JP, Waldman FM. 2004. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis* **25**: 1345–1357.
- Paoloni MC, Khanna C. 2007. Comparative oncology today. *Vet Clin North Am Small Anim Pract* **37**: 1023–1032.
- Paoloni M, Khanna C. 2008. Translation of new cancer treatments from pet dogs to humans. *Nat Rev Cancer* **8**: 147–156.
- Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci* **100**: 7672–7677.
- Prim RC. 1957. Shortest connection networks and some generalisations. *Bell Syst Tech J* **36**: 1389–1401.
- Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C. 2003. The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer* **3**: 695–701.
- Rosol TJ, Tannehill-Gregg SH, LeRoy BE, Mandl S, Contag CH. 2003. Animal models of bone metastasis. *Cancer* **97** (Suppl 3): 748–757.
- Sancho E, Batlle E, Clevers H. 2004. Signaling pathways in intestinal development and cancer. *Annu Rev Cell Dev Biol* **20**: 695–723.
- Steinbach M, Karypis G, Kumar V. 2000. A comparison of document clustering techniques. In *Proceedings of KDD Workshop on Text Mining*, pp. 109–110. ACM Press, New York.
- Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2007. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**: 90–95.
- Toft NJ, Arends MJ. 1998. DNA mismatch repair and colorectal cancer. *J Pathol* **185**: 123–129.
- Vail DM, MacEwen EG. 2000. Spontaneously occurring tumors of companion animals as models for human cancer. *Cancer Invest* **18**: 781–792.
- Yang L, Hamilton SR, Sood A, Kuwai T, Ellis L, Sanguino A, Lopez-Berestein G, Boyd DD. 2008. The previously undescribed ZKSCAN3 (ZNF306) is a novel “driver” of colorectal cancer progression. *Cancer Res* **68**: 4321–4330.
- 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 Res* **14**: 1851–1860.

Received July 24, 2009; accepted in revised form January 15, 2010.