1-1-2012

Evolutionary Dynamics of Neoplastic Cell Populations in Barrett's Esophagus

Rumen Lyubenov Kostadinov
University of Pennsylvania, rkostadi@gmail.com

Evolutionary Dynamics of Neoplastic Cell Populations in Barrett's Esophagus

Abstract

Cancer is a disease that develops over decades as result of acquisition of abnormalities in the genomes of otherwise normal cells. Acquired genomic heterogeneity in populations of cells within tissues allows cell-level Darwinian evolution that selects abnormal cellular genotypes encoding neoplastic (new benign growth), and in some cases cancerous (invasion within tissues and metastasis across tissues) cellular phenotypes. I studied neoplastic evolution over time in vivo in the pre-malignant condition Barrett's esophagus to address the puzzling clinical phenomenon that 90-95% of individuals with Barrett's stay benign over decades compared to the remaining 5-10% who progress to esophageal adenocarcinoma. Some individuals with Barrett's use aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs) that have been shown to reduce mortality from esophageal adenocarcinoma. I collaborated with the Seattle Barrett's Esophagus Research Program group to test the hypothesis that NSAIDs modulate genome evolution of neoplastic cells by reducing the acquisition rate of somatic genomic abnormalities (SGA). We used single nucleotide polymorphism (SNP) arrays to detect SGA, such as copy number abnormalities and loss of heterozygosity, in 161 biopsies from 13 individuals with Barrett's, obtained over 5-8 time points during 6-19 years of follow-up care. Over the follow-up period, each individual had a single change in NSAID use, allowing us to compare acquisition of SGA during periods on NSAIDs versus periods off NSAIDs within individuals. We found that the rate of accumulation of SGA was significantly lower (typically ten-fold lower) during periods on NSAIDs versus periods off NSAIDs. We also found that typically 1-3% of the genome had acquired SGA at baseline and that this percentage did not increase significantly over decades. In one individual who progressed to esophageal adenocarcinoma we detected a clonally expanded subpopulation of cells within the Barrett's tissue, which had massive SGA affecting 19% of the genome in the last 3 of 11 years of follow-up. In summary, these findings suggest that NSAID use may reduce SGA acquisition rate and that neoplastic cell populations in Barrett's can maintain evolutionary stasis over decades potentially explaining why 90-95% of individuals with Barrett's remain benign and never progress to esophageal adenocarcinoma.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Carlo C. Maley

Keywords

Barrett's esophagus, Cancer, Esophageal adenocarcinoma, Evolution, Genomics

Subject Categories
Genetics | Medicine and Health Sciences

EVOLUTIONARY DYNAMICS OF NEOPLASTIC CELL POPULATIONS IN BARRETT'S ESOPHAGUS

Rumen Kostadinov

A DISSERTATION
in
Genomics and Computational Biology
Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2012

Carlo C. Maley, Associate Professor, Department of Surgery,

University of California San Francisco

Dissertation Supervisor

Maja Bucan, Professor, Department of Genetics, University of Pennsylvania

Graduate Group Chairperson

Dissertation Committee:

John Maris, Professor, Division of Oncology, University of Pennsylvania

Paul Sniegowski, Professor, Department of Biology, University of Pennsylvania

Junhyong Kim, Professor, Department of Biology, University of Pennsylvania

Shane Jensen, Associate Professor, Department of Statistics, University of Pennsylvania

Mary Kuhner, Associate Professor, Department of Genome Sciences, University of Washington

EVOLUTIONARY DYNAMICS OF NEOPLASTIC CELL POPULATIONS

IN BARRETT'S ESOPHAGUS

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Carlo Maley for the kind mentorship and guidance throughout graduate school. Carlo took an evolutionary approach to cancer research problems and it has been a pleasure to work together. Thank you, Carlo! I would also like to thank my thesis committee members Dr. John Maris, Dr. Junhyong Kim, Dr. Mary Kuhner, Dr. Paul Sniegowski, and Dr. Shane Jensen for their selfless mentoring when I needed advice in cancer biology, evolutionary biology, computation and statistics, and scientific writing. I especially thank Mary Kuhner for advising me weekly during the last 6 years.

I thank all my labmates Lauren Merlo, Kathleen Sprouffske, Rachelle Kosoff, Aleah Fox Caulin, Kristin Gardiner, and Trevor Graham for all their support and for the countless scientific discussions. In particular, a number of tricky problems have been made clearer through discussions with Kathleen Sprouffske. The last two years as a guest in Junhyong Kim's lab have been enjoyable and I thank all the Kim lab members for being kind and helpful.

I am indebted to all my collaborators from the Seattle Barrett's Esophagus Program team. I am very thankful to Dr. Brian Reid for his insightful mentorship. My studies would not have been possible without the help of Rissa Sanchez, Patty Galipeau, Thomas Paulson, Xiaohong Li, Patty Blount, David Cowan, and Thomas Vaughan, and I thank them for their kind support and advice.

I thank the Genomics and Computational Biology faculty, from whom I learned a lot; in particular I am grateful to Warren Ewens, Harold Riethman, Sridhar Hannenhalli, and Maja Bucan. The students and staff of my graduate group have been a great support for me, in particular I would like to thank Hannah Chervitz, Logan Everett, Paul Ryvkin, Sameer Soi, Nick Stong, and Hannah Dueck.

Most importantly, I would like to thank my parents Lyuben and Yanka Doychevi, and my sister Kremena Kostadinova, for their loving support and encouragement throughout graduate school.

ABSTRACT

EVOLUTIONARY DYNAMICS OF NEOPLASTIC CELL POPULATIONS IN BARRETT'S ESOPHAGUS

Rumen Kostadinov

Carlo Maley

Cancer is a disease that develops over decades as result of acquisition of abnormalities in the genomes of otherwise normal cells. Acquired genomic heterogeneity in populations of cells within tissues allows cell-level Darwinian evolution that selects abnormal cellular genotypes encoding neoplastic (new benign growth), and in some cases cancerous (invasion within tissues and metastasis across tissues) cellular phenotypes. I studied neoplastic evolution over time *in vivo* in the pre-malignant condition Barrett's esophagus to address the puzzling clinical phenomenon that 90-95% of individuals with Barrett's stay benign over decades compared to the remaining 5-10% who progress to esophageal adenocarcinoma. Some individuals with Barrett's use aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs) that have been shown to reduce mortality from esophageal adenocarcinoma. I collaborated with the Seattle Barrett's Esophagus Research Program group to test the hypothesis that NSAIDs modulate genome evolution of neoplastic cells by reducing the acquisition rate of somatic genomic abnormalities (SGA). We used single nucleotide polymorphism (SNP) arrays to detect SGA, such as copy number abnormalities and loss of heterozygosity, in 161 biopsies from 13 individuals with Barrett's, obtained over 5-8 time points during 6-19 years of follow-up care. Over the follow-up period, each individual had a single change in NSAID use, allowing us to compare acquisition of SGA during periods on

NSAIDs versus periods off NSAIDs within individuals. We found that the rate of accumulation of SGA was significantly lower (typically ten-fold lower) during periods on NSAIDs versus periods off NSAIDs. We also found that typically 1-3% of the genome had acquired SGA at baseline and that this percentage did not increase significantly over decades. In one individual who progressed to esophageal adenocarcinoma we detected a clonally expanded subpopulation of cells within the Barrett's tissue, which had massive SGA affecting 19% of the genome in the last 3 of 11 years of follow-up. In summary, these findings suggest that NSAID use may reduce SGA acquisition rate and that neoplastic cell populations in Barrett's can maintain evolutionary stasis over decades potentially explaining why 90-95% of individuals with Barrett's remain benign and never progress to esophageal adenocarcinoma.

TABLE OF CONTENTS

Chapter 1. Introduction

1.1. Progression of Barrett's esophagus to esophageal adenocarcinoma

This doctoral dissertation relates to a pressing human need – to cure cancer. Cancer is a disease that can end a human life prematurely. A research study by the American Cancer Society estimates that, in the year 2012, 1,638,910 people would be diagnosed with cancer and 577,190 people would die from cancer in the U.S., or stated simply, nearly one of every four deaths in the U.S. is associated with cancer [1]. **In this work, I investigate the phenomenon of within-individual change in the genome of neoplastic cell populations over time and I investigate and suggest ways to control somatic genomic change over time to prevent cancer and extend human health span.** I hypothesize that the genomes of normal and neoplastic cells acquire somatic genomic abnormalities gradually over time resulting in progressive transformation to invasiveness and malignancy. Throughout this work, I examine the somatic genomic change over time in neoplastic cell populations through evolutionary lens by adhering to clonal evolutionary theory [2] according to which neoplastic cell populations evolve by acquisition of somatic genomic change permitting Darwinian natural selection of variant cell lineages. I hypothesize that measuring somatic genomic evolutionary dynamics of neoplastic cell populations within individuals with Barrett's esophagus would help distinguish individuals at low and high risk of progression to esophageal adenocarcinoma that would help prevent overdiagnosis and overtreatment in low-risk individuals and underdiagnosis and undertreatment in high-risk individuals.

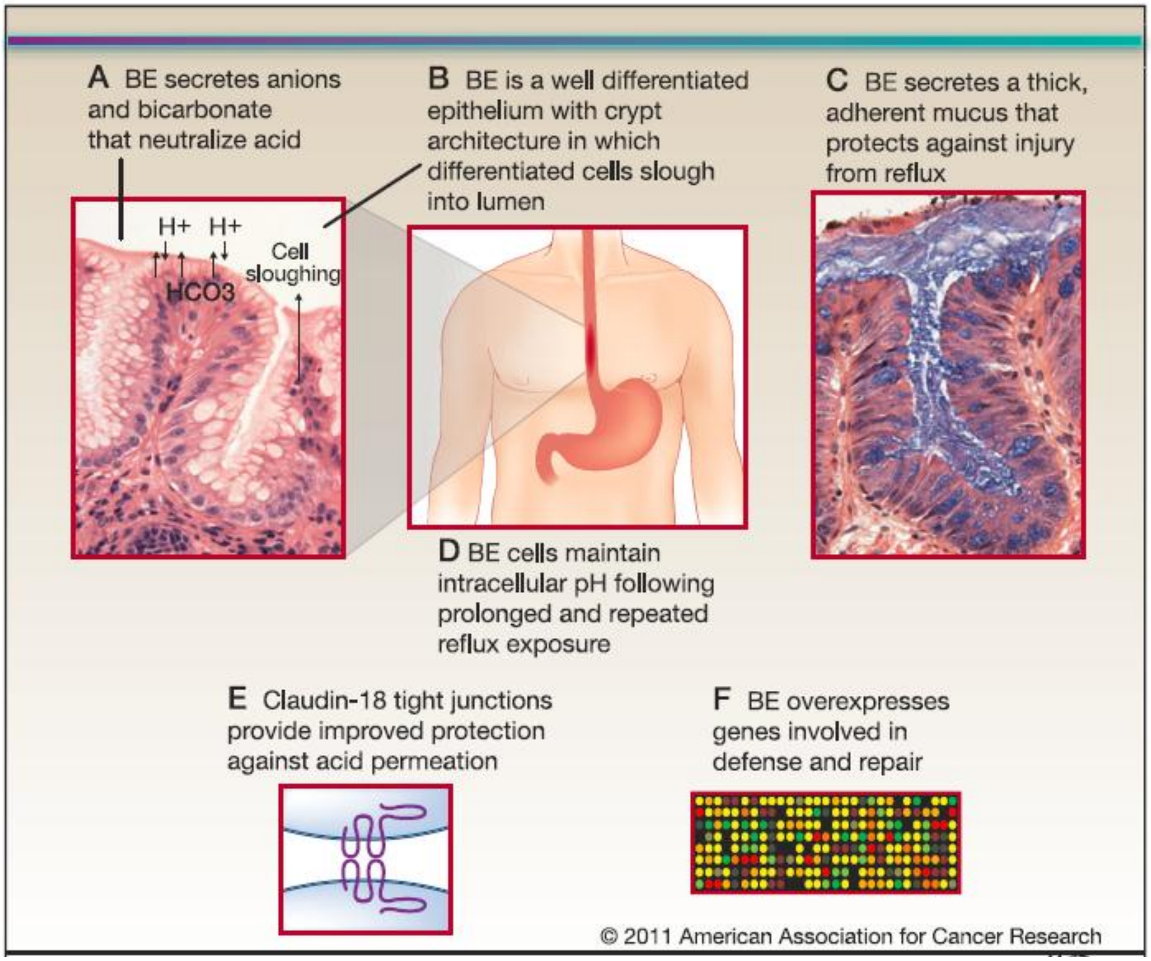
Does carcinogenesis have clearly defined stages? Berenblum first described quantitatively in mouse that the process by which cells become hyperplastic, that is, acquiring overgrowth compared to neighboring normal cells, can be divided into two stages, "initiation" and "promotion", where cells can be initiated by application of a carcinogen, but they may lay latent or dormant within the tissue, unless a distinct promoting agent triggers clonal expansion manifesting hyperplasia [3]. Moreover, Berenblum demonstrated that distinct chemical carcinogens vary in their ability initiating and promoting abilities [3]. Individuals with Barrett's esophagus enter clinical management after being symptomatic, and biopsies are collected only if a Barrett's segment is already present, thus we can never observe the *in vivo* initiation of Barrett's neoplasia and can only infer the elapsed time since initiation indirectly based on initiation hypotheses and modeling assumptions. Foulds first described the term "progression", in the context of chemical carcinogenesis in model organisms, where progression can be thought of as a distinct stage in neoplastic development marked by karyotypic alterations in cell genotype and increased cellular growth rate, invasiveness, and metastasis in cell phenotype [4]. My studies are limited to observing the genomic changes of already hyperplastic and neoplastic cells over time, or the promotion, and in only one case the promotion and progression phases in neoplastic evolution in Barrett's esophagus and its progression to esophageal adenocarcinoma.

My doctoral thesis seeks to shift the target of interventions from targeting individual genes to targeting the dynamics of the whole genome. By characterizing the whole genome I aim towards a holistic approach to the problem by measuring evolutionary biology parameters, such as mutation rate, clonal expansion rate, magnitude of genetic diversity over time, phylogenetic tree imbalance, etc., and associating such measures with clinical variables as opposed to aiming towards a reductionist approach by associating abnormalities in individual genes with clinical variables. Can we prevent some cancers by lowering the rate of genomic changes or the clonal expansion rate of genetically unstable clones? The field is ripe for such a change because, although somatic evolution is widely accepted as the theory of cancer, the tools from evolutionary biology for studying that process have not yet been widely adapted to study and prevent cancer. In my view, a theory of cancer would not be complete without pairing theoretical models of genome dynamics over time with experimental *in vivo* observations of genome dynamics over time in human biopsy specimens.

1.2. Barrett's esophagus biology

Barrett's esophagus (BE) is a condition of the distal esophagus in which the normal stratified squamous epithelium is replaced by columnar epithelium with intestinal metaplasia [5]. BE is thought to develop as complication of chronic gastroesophageal reflux disease (GERD) and individuals with BE are at increased risk of progression to esophageal adenocarcinoma (EA): 1-7 persons with BE progress to EA per 1000 person-years [6,7]. Scientific progress is providing new understanding of previously puzzling phenomena. **The puzzling phenomenon in Barrett's esophagus is why 90-95% of individuals with BE follow a benign course and never develop EA, and accordingly why the remaining 5-10% of individuals do progress to EA.** I investigate this phenomenon by investigating the whole genome of Barrett's neoplastic cells from biopsies collected from 13 individuals over 6-19 years, with the hypothesis that the change in the number of acquired somatic genomic alterations over time may provide an insight into the phenomena of retaining a benign course over decades versus developing malignancy.

Reid, Kostadinov, and Maley suggested new strategies for clinical management of Barrett's esophagus by integrating clonal evolutionary theory into clinical diagnosis and practice [8]. The current and novel paradigms in the biology of Barrett's metaplasia are summarized in Figure.


A BE secretes anions and bicarbonate that neutralize acid
H+ H+
HCO3
Cell sloughing
B BE is a well differentiated epithelium with crypt architecture in which differentiated cells slough into lumen
C BE secretes a thick, adherent mucus that protects against injury from reflux
D BE cells maintain intracellular pH following prolonged and repeated reflux exposure
E Claudin-18 tight junctions provide improved protection against acid permeation
F BE overexpresses genes involved in defense and repair

© 2011 American Association for Cancer Research

Figure 1.1. Barrett's specialized intestinal metaplasia and mucosal defense. Barrett's metaplasia arises in an environment of chronic reflux in which the distal esophagus is exposed to high levels of local and systemic damage from acid, bile, and tobacco products, as well as inflammatory responses to the injury [5,9–14]. All are mutagenic. Barrett's metaplasia has a number of defenses against this mutagenic environment that are not found in esophageal squamous epithelium [5,15]. A, Barrett's metaplasia secretes anions, including bicarbonate, that participate in buffering acid reflux [16]. B, Barrett's metaplasia is a well differentiated epithelium with crypt architecture in which putative stem cells residing at the base give rise to proliferating transient amplifying cells and differentiated cells that are

sloughed into the lumen. This architecture has been proposed to be tumor-suppressive because mutations in transient amplifying or differentiated non-stem cells will be shed from the body before they can accumulate the serial mutations that lead to cancer [17]. C, Barrett's metaplasia secretes a thick adherent mucus that is not present in squamous esophageal epithelium for defense against acid and bile reflux [18–21]. D, Barrett's esophageal cells maintain physiological intracellular pH after prolonged and repeated reflux exposure [22]. E, The tight junctions of Barrett's metaplasia overexpress claudin 18 and several other claudins (including claudins 1, 4, 12, and 23) that provide protection against acid permeation [23]. F, A combined expression and proteomics study of Barrett's metaplasia reported overexpression of genes involved in mucosal defense and repair [24]. Figure and figure legend adapted from Reid et.al. [8].

The British oncologist Willis defined the essence of a neoplasm as "A neoplasm … the growth of which persists in the same excessive manner after cessation of the stimuli which evoked the change" [25]. The persistence of Barrett's metaplasia in the esophagi of BE individuals fits this description of a neoplasm, where whichever stimuli first evoked the replacement of multi-layer squamous epithelium to single-layer columnar epithelium (metaplasia) the cessation of such stimuli, by acid suppression medications or anti-reflux surgery do not appear to result in regression of the Barrett's segment. I hypothesize that the persistence of Barrett's metaplasia can be explained by irreversible genomic changes, where a genome that has lost genomic information by deletion and loss of heterozygosity, cannot regain such information back to revert back to expressing normal function. I evaluated the length of the Barrett's segment length over decades of follow-up time in 248 individuals with Barrett's participating in the Seattle Barrett's Esophagus Research Program. I found that the segment length remains 5.7 ± 3.6cm (mean ± standard deviation) long over decades (Figure 1.2) that shows both the persistence of this tissue despite acid suppressive medications, and the relative constancy in the neoplastic cell population size of a Barrett's segment.

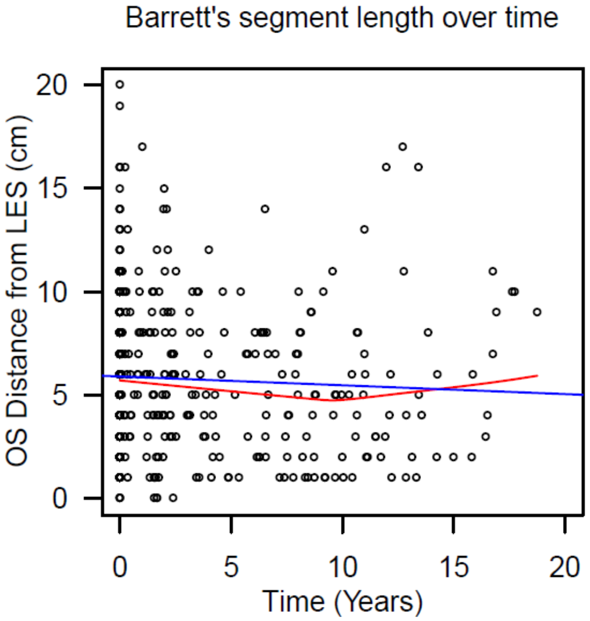


Figure 1.2. Typical Barrett's segment length remains about 5cm during endoscopic surveillance. Segment length data over time represents 248 individuals with BE from the Seattle Barrett's Esophagus Program cohort. Each point represents an estimate of the distance between measured lower esophageal sphincter (LES) and *ora serrata* (OS) that are recorded at an endoscopy (out of 489 total) and that are proxies for the extent of the Barrett's segment. Over all 489 endoscopies, the typical Barrett's segment length measures 5.7 ± 3.6cm (mean ± s.d) and linear (blue line) and local weighted linear regressions (red curve) show that segment lengths stay relatively constant over time.

The following chapters explore the genomic changes and evolutionary dynamics of neoplastic cells occurring within BE tissues having this apparent constancy in length over time. The word neoplasm is concisely defined as "tumor; any new and abnormal growth, specifically one in which cell multiplication is uncontrolled and progressive" and the word cancer is concisely defined as "a neoplastic disease the natural course of which is fatal" [26]. These two definitions capture the fundamental biology of cancer as a disease, namely the phenomenon of uncontrolled and progressive overgrowth of cell populations that may have a fatal outcome. These two definitions also present the crux of what I define as **the cancer diagnosis problem: Can we diagnose neoplasms (tumors) and predict which ones**

would develop into malignant cancers that have a fatal natural course and which ones would instead stay benign and not be fatal?

1.3. The application of evolutionary theory to cancer

While the Dorland medical dictionary refers to the cancer disease as having a "natural course", I conceptualize the cancer disease as having an "evolution", that is, a somatic cell level theory of Darwinian evolution, or viewing the cancer disease development as an evolutionary and ecological process [27]. Charles Darwin put forward that the mechanism of evolution is natural selection, which has three requirements: 1) variation in a population of individuals 2) such variation must be heritable, and 3) such heritable variation must confer differential fitness, i.e. a heritable trait that confers better reproductive or survival advantage to individuals would increase in frequency in the population over several generations. However, Darwin avoided the term evolution and instead preferred to use the term "descent with modification" possibly because evolution implied a pre-charted trajectory or a premeditated point to be reached by an evolving biological species. Instead of defining neoplastic evolution as "a process of change in a certain direction", I limit neoplastic evolution to also mean "descent with modification" implying that the process itself is inherently random and no premeditated endpoint exists and no trajectory of evolution's course can be predicted. In other words, neoplastic evolution in Barrett's esophagus does not necessarily have to have a premeditated adenocarcinoma endpoint. In cancer, "descent with modification" describes the process by which somatic cell populations evolve or descend with modification by the mechanism of natural selection. Natural selection applies to somatic cell populations since they can: 1) acquire somatic genomic abnormalities (SGAs), such as DNA point mutations, copy number alterations, and structural chromosomal rearrangements, 2) such SGAs are heritable over generations since cells divide asexually by mitosis that copies SGAs from a parent cell's DNA to a daughter cell's DNA, and 3) such heritable SGAs can confer differential fitness to cells that bear them, i.e. survival or reproductive advantages manifested by abnormal overproliferative somatic cell phenotypes. Acquisition of SGAs alters the genomes of normal somatic cells within somatic tissues and by the mechanism of natural selection normal somatic cell populations can evolve hyperplasia, neoplasia, and malignancy by selfishly increasing in numbers, or increasing fitness relative to neighboring normal cells. Neoplastic cell populations that acquire SGA begin exhibiting one or more of the cell phenotypes ("hallmarks") of cancer by altering the molecular pathway networks of somatic cell homeostatic cell growth, cell death, and cell designated tissue function [28].

1.4. The theory of clonal evolution, or the theory of the genetics of neoplastic development

To begin with, I will define some terms from my point of view. The neo-Darwinian synthesis brings together Darwin's theory of descent with modification and Mendelian genetics. Population genetics is a field of study that rapidly advanced since the neo-Darwinian synthesis and shifted focus from individual thinking to population thinking. Mathematical population genetics provides a theoretical framework for describing the changes in the genetic constitution of populations over time, most notably, changes in allele frequencies over time. Phylogenetics can be categorized as a subfield of population genetics, which is primarily concerned with estimating the morphologic (phenotypic) or genetic (genotypic) relatedness among species in a population of species or among individuals in a population of individuals. Evolutionary dynamics is the process of change over time in a population of individual entities that can evolve, i.e. the change in certain characteristics of the population as result of the processes of mutation, natural selection, and migration (expansion in space) over time within the population.

Several scientific advancements put forward evolutionary thinking as it relates to neoplastic development. Theodor Boveri first theorized that SGA may underlie neoplastic growth by observing in frog embryos multipolar mitoses that caused improper segregation of DNA to daughter cells and resulted in the generation of genetic variation and phenotypic abnormalities [29]. Later, in early cytogenetic studies Peter Nowell discovered the Philadelphia chromosome, a common chromosomal rearrangement in chronic myeloid leukemias and suggested in 1976 that "the acquired genetic variability permits stepwise selection of variant cell lineages that underlies tumor progression", which became known as the clonal evolution theory of neoplastic progression [2]. In 2006, Merlo et al. reviewed multiple studies to support the notion that neoplastic progression is an evolutionary and an ecological process [27] suggesting that clonal evolution theory has survived challenges for more than 40 years and can be thought of as the underlying theory of cancer.

Tumor genetic heterogeneity and aberrant karyotype (aneuploidy) is a ubiquitous observation in primary tumors and metastatic cancers. Maley et al. showed that increased genetic diversity in BE is associated with increased risk of progression to EA [30]. Merlo reviewed the role of measuring genetic diversity in cancer [31]. Two studies by Park et al. showed that increased genetic heterogeneity in breast cancer was associated with poor clinical outcomes [32,33]. Leedham et al. showed individual crypt heterogeneity within Barrett's segments by using LOH markers and p16 and p53 sequence data [34]. A recent study by Beroukhim et al. of 3131 cancer specimens from 26 histologic types showed widespread somatic genomic alterations (SGA) involving both focal and chromosome-wide regions of the genome: 75,700 copy number gains and 55,101 copy number losses were identified, covering an average of 17% and 16% of the genomes, respectively, in cancer samples, compared to an average of 0.35% and 0.1% in normal samples [35]. Vogelstein and colleagues advanced a chromosomal instability theory underlying cancer development [36–40]. Despite the recent massive accumulation of data on genetic alterations in cancers, the process of somatic genomic evolution that leads to cancer is severely understudied, even though it is the driving force for neoplastic progression. The challenge for cancer prevention is to halt or delay the genome instability that leads to somatic genomic evolution and progression to cancer.

1.5. Application of phylogenetic tree estimation or cell lineage reconstruction in cancer data

Since mitotic cell division results in one cell replicating in two, assuming no horizontal gene transfer, the cell lineage history of a mitotically dividing cell population must have a phylogenetic tree structure. Navin et al. recently reviewed the potential of tracing individual cell lineages in tumors [41]. Also, Navin et al. emphasize the utility of sectioning a primary tumor into sectors thereby preserving spatial locations of individual sectors, and evaluating sectors by flow cytometry or by array-CGH for genomic abnormalities in order to infer the natural history of tumor progression [42]. Salipante and Horwitz review the utility of capturing genomic variation at genomic sites that have high enough somatic mutation rate in order to infer the phylogenetic fate of cell lineage [43]. Frumkin et al. and Wasserstrom et al.

also show that parallel measurements of microsatellites can be used for inferring cell lineages of cells in various tissues within human or mouse [44,45]. Frumkin e. al. also showed the feasibility of such approach by typing microsatellites of a mouse tumor and its associated metastases to reconstruct the temporal order of tumor progression [46]. Both Horwitz and Shapiro groups develop their own technologies for typing multiple distinct sets of microsatellites, usually using capillary electrophoresis and robotics for parallelization throughput. However, those technologies quickly become outdated by SNP genotyping and next-gen sequencing of entire genomes at ever smaller scales: from micrograms of DNA to DNA of single cells. A 2012 review from Carlson et al. from Horwitz's group confirms the utility of deep sequencing technologies for achieving ever more accurate cell fate maps [47].

One of the first demonstrations of tree estimation was in Louhelainen et al. who typed LOH at 87 loci of multiple biopsies from multifocal bladder cancers and found that multiple biopsies appeared to share similar LOH abnormalities thereby suggesting monoclonal origin as opposed to independent multiclonal origins of those cancers [48]. Ruiz et al. used flow cytometry, FISH, and PCR to show clonal evolution in a single individual with prostate adenocarcinoma by analyzing serially sampled (longitudinal) biopsies [49]. Clonal analysis from 40 pancreatic adenocarcinoma in the same study mapped genomic alterations in primary and metastatic sites within individuals. The authors concluded that variant clonal populations can exist within and between biopsies of the same primary tumor and characterization of the genome and clonal analysis can help manage treatment options and evaluate the evolutionary response of treatments. However, they did not estimate clonal relationships using evolutionary methods, such as phylogenetic tree estimation, based on the shared patterns of DNA content abnormalities and the shared patterns of FISH probe presence and absence. Salk et al. showed that poly-guanine tract (a mononucleotide repeat microsatellite locus, i.e. a string of "G"s) loci can be useful neutral markers to identify fields of clonal expansion in patients with ulcerative colitis [50]. Salk et al. used PCR of poly-G tracts to show that detection of increased clonal expansion increased risk of progression to colorectal cancer [51]. Tsao et al. sampled mismatch repair deficient (MMR-) colorectal tumors and measured the length of microsatellite loci with PCR to estimate the mitotic age of tumors based on a mathematical model of neoplastic progression, based on the expectation of increased variation in microsatellite length as the tumor undergoes successive mitoses over time [52]. Campbell et al. detected multiple co-existing subclonal cell populations within individual patients (n=22 patients total) with B-cell chronic lymphocytic leukemia using ultra-deep pyrosequencing of the immunoglobulin locus, since rearrangement at that locus occurs at high frequency, i.e. the somatic mutation rate at that locus is high making it a useful neutral marker of clonal evolution [53]. In that study, an unrooted parsimony in PHYLIP was used on sequences that were multiply aligned with CLUSTALW2. Yachida et al. inferred a pathway series of sequence alterations in pancreatic adenocarcinoma [54]. All the above studies exemplify the application of evolutionary approaches to cancer data, i.e. estimation of phylogenetic relatedness among neoplastic and/or cancer cells.

1.6. Modeling studies of genetic diversity and clonal evolution

The acquisition of somatic genomic abnormalities results in heterogeneous neoplasms and computer modeling of genomic diversity in neoplasms can help understand plausible scenarios of how the rates of clonal lineages growth, spread, and mutation affect the overall spatial genetic structure of the entire cell population. Chao et al. used agent-based modeling and experimental skin lesion experimental data and found that the speed of clonal expansion is driven by death of neighboring cells and appears to follow quadratic rather than exponential growth [55]. Martens et al., in collaboration with Carlo Maley and myself, used a 2D spatial stochastic Moran model and analytical formulations to show that most parameter conditions may result in clonal expansions that arise independently at distinct locations within a neoplasm and come into contact, or interfere, with each other's expansion, which results in increased waiting time to cancer [56]. This effect is also termed Hill-Robertson (clonal) interference. Only small neoplasm size and strong selection will favor periodic selection over clonal interference. Graham et al. hypothesized three plausible scenarios that generate the high genetic diversity that is commonly observed in Barrett's esophagus: high mutation rate coupled with strong selection, synergizing mutually beneficial interactions between heterogeneous clones, and an unidentified landscaping alteration driving an initial clonal expansion on the back of which neutral mutations can hitchhike generating diversity [57]. Sottoriva et al. developed an agent-based model to explore the generation of epigenetic methylation patterns during clonal expansions of a growing neoplasm [58]. Nicolas et al. used crypt methylation patterns, generated by bisulfite treatment and sequencing the *BGN* locus on chromosome X, to infer population parameters, such as stem cell number per crypt, in colorectal cancer [59]. This pioneered the use of approximate Bayesian computation for inference of parameters by matching summary statistics from observed experimental data to summary statistics obtained from simulated data. While the above recent advancements in modeling have illuminated aspects of genetic diversity dynamics, more modeling is needed in matching models to observations in specific systems, such as Barrett's esophagus, in order to bound some of the parameter space for explanatory power and for specific predictions for currently unobservable underlying dynamics.

1.7. Advancements in backward-in-time estimation of the genetic history of a population

Recent advancements in computing power allowed the application of Kingman's coalescent theory (for history of coalescent theory refer to [60]) to genetic data, for estimation of population genetic parameters. Kuhner et al. developed Bayesian Markov Chain Monte Carlo (MCMC) methods for estimating population genetic parameters from SNP data [61]. Drummond et al. developed the software BEAST, which is a flexible Bayesian inference framework for population-genetic parameter inference from serially sampled data [62]. Excoffier et al. developed the software SerialSimCoal, which provides a forward simulation of population demographics and mutation under various scenarios and parameters, importantly producing temporally spaced samples [63]. Both BEAST and SerialSimCoal software programs were, and have been to date, open-source that allowed me to modify their code and add on novel evolutionary models for novel types of genetic data, such as data for copy number and LOH abnormalities from SNP array data.

1.8. Chemoprevention in Barrett's esophagus to prevent progression to esophageal adenocarcinoma

NSAIDs have a strong and significant effect on reducing the incidence and mortality of many cancers, including esophageal, colorectal, lung, and other malignancies, [Hazard ratio was 0.66 (95% CI 0.50-0.87) for all malignancies], however their effect manifests significantly after more than 5 years of regular use [64]. The majority of epidemiological evidence suggests that NSAID use in individuals with BE reduces risk of developing EA [65–68]. Vaughan et al. evaluated 350 individuals followed up for a median of 5.4 years (range 0.2-8.9) and showed that the 5-year cumulative incidence of EA was 14.3% (95% CI 9.3-21.6) for never users and 6.6% (3.1-13.6) for current NSAID users [67]. The hazard ratio for EA for NSAID users was 0.20 (95% CI 0.10-0.41) compared with NSAID non-users [67]. Galipeau et al. showed that NSAID use modulates the risk of developing DNA content abnormalities (tetraploidy and/or aneuploidy), assayed by flow cytometry, and genetic abnormalities, such as loss of heterozygosity (LOH) on chromosomes 9p and 17p, assayed by PCR of small tandem repeat (STR) loci [68]. The benefits of NSAID use for chemoprevention are attractive due to their widespread use and low toxicity; however the molecular mechanisms underlying their preventive effect are not fully understood. I evaluate 13 individuals with Barrett's in Chapter 4 to show that NSAIDs modulate acquisition of SGA in neoplastic cell populations.

1.9. Conclusion

In summary, in the next chapters, the population genomic transformation of neoplastic cells in Barrett's esophagus would be evaluated *in vivo* in human over decades of follow-up. Hopefully, the reader would gain novel insights into the simple central question: "How does the somatic genome evolve over time in the context of neoplastic evolution and progression to cancer?", as I have attempted to best describe qualitatively and quantitatively using novel technologies for genomic characterization and novel application of methods from evolutionary biology.

Chapter 2. Pilot studies evaluating genomic DNA of Barrett's esophagus biopsies for somatic genomic abnormalities

2.1. Chapter Introduction

Ever since the sequencing of the human genome [69,70], rapidly emerging commercial platforms and technologies allowed evaluating the DNA of human tissue specimens for copy number and loss of heterozygosity abnormalities. At this time, 200 nanograms of DNA extracted from a human biopsy sample can be evaluated with several commercially available technologies that have probes to evaluate about 1 to 2 million locations of the 3.2 billion base pair human genome for about four hundred dollars per sample. Full sequencing of entire genomes is also available, though the cost is about five thousand dollars per sample (about a microgram of DNA). Throughout my studies I worked with data from six platforms for genome-wide DNA assessment: Illumina 33K, Illumina 109K, Illumina 317K, Illumina 550K, Illumina 1M, and Affymetrix 1M (SNP6.0), single nucleotide polymorphism (SNP) arrays. I analyzed data from several pilot studies that were designed to evaluate the capabilities of various technologies to detect accurately and reproducibly DNA copy number changes and loss of heterozygosity changes in genomic DNA from human tissue specimens.

I collaborated with the Seattle Barrett's Esophagus Program (SBEP) group to design and execute several pilot studies to evaluate the Illumina platform, which provided preliminary evidence for larger scale (case control and case cohort) studies of SGA in biopsies from individuals with Barrett's in the SBEP cohort.

The Illumina SNP platform is based on magnetic bead chips designed to detect single nucleotide polymorphisms (SNPs) in a human sample. The main application of SNP array technology has been to call out the major or minor allele (nucleotide variants at a single base pair position on a chromosome) at each of 33,000 (Illumina Infinium 33K array) to 1 million (Illumina OmniQuad 1M array) positions in the human genome. The design and manufacturing of a SNP array can influence significantly the quality of results, usually summarized with two metrics: the signal to noise ratio and the SNP call rate. Gunderson et al. showed that the advantage of Illumina's SNP arrays in terms of reducing spatial biases across arrays is the barcoding of magnetic beads and the random spread of beads on a glass slide hexagonal lattice during manufacturing and scanning of the barcodes and SNP calling during single-base pair extension reaction emitting a fluorophore [71]. Peiffer et al. showed the utility of Illumina SNP arrays for detecting regions of DNA copy number alterations and loss of heterozygosity [72]. The developments of the Illumina SNP platform increased density of the probes over time allowing an increased genomic coverage to detect break points of genomic abnormalities at a higher resolution and to detect small indels.

The SBEP group had evaluated Illumina 33K arrays prior to 2006 and I started a collaboration with them to evaluate Illumina 109K arrays in 2006, Illumina 317K arrays in 2006-2009, and Affymetrix SNP6.0 and Illumina OmniQuad 1M arrays in 2009-2012. We performed several pilot experiments, four of which I describe below, to assess somatic genomic abnormalities (SGA) in Barrett's biopsies.

2.2. Pilot experiment evaluating Illumina 109K SNP platform

2.2.1. Introduction and Methods

One of the first pilot experiments we performed was evaluating the signal to noise ratio in Illumina 109K-SNP arrays. We selected two biopsy samples, two gastric samples, and one lymphocyte sample from the same BE individual. The lymphocyte sample contains leukocytes (white blood cells) that contain DNA that represent the germline genotype of the individual, that is, the DNA lacks any somatic genomic abnormalities. Usually, the DNA from a gastric biopsy also lacks somatic genomic abnormalities. The aim of the experiment was to test signal quality difference between two halves of the same biopsy and between two biopsies from the same patient. The biopsy samples were processed at FHCRC (for full method description of sample processing refer to Methods in Chapter 4). The final sample sheet and raw array data were provided to me.

Sample ID	Sample type	Input DNA (ng)
2	Gastric	60
3	Gastric	30
4	BE1	30
5	BE1r (BE1 replicate)	45
6	BE2	30
7	Lymphocyte	150
8	BE2r (BE2 replicate)	75

Table 2.1. Pilot study design for evaluating SGA with Illumina 109K SNP array technology.

I processed raw SNP array data with Illumina's BeadStudio software, and wrote Perl algorithms to detect loss of heterozygosity (LOH) based on genotype calls output from BeadStudio. I defined a SNP as "LOH-informative" if a SNP is heterozygous ("AB") in leukocyte or gastric sample, representing the germline state of the SNP, since such SNPs would be informative for detecting loss of heterozygosity when such SNPs get an "AA" or "BB" genotype call in Barrett's epithelium samples. I developed a Perl algorithm to iterate trough every SNP along a chromosome and define contiguous regions of LOH (blocks of LOH), such that contiguous regions contained more than one LOH-informative SNPs. Also, the algorithm was parameterized to trust the call of a single LOH-informative SNP (version A) or to trust the call of two contiguous LOH-informative SNPs (version B). In other words, in version B, if any one of two neighboring LOH-informative SNPs is not called LOH, then both SNPs get a non-LOH call, which effectively eliminates incorrectly called singlet SNPs, and increases LOH region calls stringency. In both versions A and B, two neighboring LOH loci could be separated by LOH-uninformative loci, which are discarded from consideration. Genotype call concordance was defined as the proportion of SNPs sharing the same genotype call between two samples, out of a total of 109,365 SNPs.

2.2.2. Results and Discussion

The total number of SNP probes on the 109K SNP platform was 109,365 and when comparing the genotype calls between samples 2, 3, and 7, there was 99.27% genotype call agreement among all of the three samples, or only 791 SNPs had discordant calls between any of the three control samples. This suggests both that it is unlikely that any of these samples contain a DNA abnormality and that 99.27% of the SNP probes gave out signal intensity that after normalization and processing resulted in accurate genotype calls. Also, in this individual, 35,788 SNPs had a heterozygous "AB" call, thereby labeling 35,788 SNPs as "LOH-informative" SNPs, and 72,786 SNPs had a homozygous "AA" or "BB" call, thereby discarding them from LOH analysis. As opposed to apparently normal genotype in samples 2,3, and 7, hereafter referred to as the control samples, multiple somatic genomic abnormalities were detected in the BE samples.

A total of 719 LOH-informative SNPs showed LOH in any of the samples 4, 5, 6, and 8 when using version A of the LOH detection algorithm. The concordance when comparing the same biopsy split in half and run on two separate arrays (within biopsy concordance) was 72.11% and 78.28%. The concordance when comparing two biopsies collected from two different levels from the same endoscopy from the same individual (between biopsy concordance) ranged from 60.49% to 63.57%.

n=719	**BE1r**	**BE2**	**BE2r**
BE1	72.11%	63.21%	63.57%
BE1r		60.49%	62.82%
BE2			78.28%

Table 2.2. Genotype call concordance within and between biopsies by trusting genotype calls of each LOH-informative SNP when detecting contiguous LOH regions. Samples labeled with "r" are the same biopsy split in half, whereas samples labeled with "1" and "2" are biopsies collected from two different levels in the Barrett's segment.

n=272	**BE1r**	**BE2**	**BE2r**
BE1	98.75%	82.89%	82.15%
BE1r		82.65%	81.25%
BE2			96.88%

Table 2.3. Genotype call concordance within and between biopsies by trusting genotype calls of two neighboring LOH-informative SNPs when detecting LOH regions.

Because of apparently low concordance in SNP calls (Table 2.2) we used version B to call regions of LOH by trusting the genotype call of two neighboring LOH-informative SNPs when evaluating LOH event break points. A total of 272 LOH-informative SNPs showed

LOH in any of the samples 4, 5, 6, and 8 (Table 2.3). The concordance when comparing the same biopsy split in half and run on two separate arrays (within biopsy concordance) was 98.75% and 96.88% (Table 2.3). The concordance when comparing two biopsies collected from two different levels from the same endoscopy from the same individual (between biopsy concordances) ranged from 81.25% to 82.89% (Table 2.3).

2.2.3. Conclusion

The Illumina 109K SNP platform performed well for identifying regions of LOH from isolated DNA from BE biopsies. The results showed that DNA isolated from lymphocyte samples and gastric tissue specimens appear normal, that is, lacking any DNA LOH abnormalities. When biopsies were split in half and evaluated with two separate SNP arrays, SNPs in regions of LOH showed 72-78% concordance in genotype calls, suggesting very low between-array variation in signal quality. However, we noticed that the genotype call of a single SNP cannot be trusted and when we increased stringency by trusting the combined call of two neighboring LOH-informative SNPs, the concordance in genotype calls increased to 97-99%, suggesting that bioinformatic methods can improve stringency and accuracy of LOH calls.

2.3. Pilot experiment evaluating Illumina 317K SNP platform: clonal evolution in one individual with Barrett's esophagus over 16 years of follow-up

2.3.1. Introduction and Methods

We designed a longitudinal study of a single individual with Barrett's esophagus to test performance of the Illumina 317k platform and to continue developing computational methods for processing Illumina SNP data. We selected an individual who had been in endoscopic surveillance for over 16 years and the morphology diagnosis had been metaplasia throughout follow-up endoscopic surveillance. We had no specific prior expectation of the dynamics of somatic genomic abnormalities (SGA) and we only hypothesized that we would observe accumulation of SGA over time. The aim of this study was to develop computational methods for handing longitudinal data from BE biopsies from the Illumina platform and develop methods for estimating phylogenies and estimating population-genetic metrics, such as rate of acquisition of SGA.

id #	**date**	**biopsy Level**	**sample type**	**ng**
1	1989	33	BE	49
2	1989	29	BE	205
3	1989	29	BE	70
4	1993	32	BE	112
5	1993	30	BE	34
6	1993	26	BE	71
7	2001	36	BE	120
8	2001	32	BE	182
9	2001	30	BE	94
10	2006	30	BE	118
11	2006	28	BE	191
12	2006	26	BE	202
13	2006	N/A	blood	104
14	2006	N/A	blood	35
15	2006	43	gastric	200
16	2006	50	gastric	200

Table 2.4. Sixteen longitudinal samples from the same individual were analyzed with Illumina 317K-SNP arrays. DNA from Barrett's samples was extracted using epithelial isolation technique [73] (for detailed sample preparation information see Methods in Chapter 4). DNA from blood and gastric samples were extracted without the epithelial isolation step. Biopsies were taken from various time points and various levels from the Barrett's segment. Various final amounts of DNA were used on the arrays to compare signal-to-noise ratio correlation with input DNA amount.

All raw intensity files were loaded in Illumina's BeadStudio, normalized and clustered using the SNP manifest and canonical genotype cluster files, provided by Illumina, for build36 of the human genome. In the following analyses we used the normalized, total signal intensity "R" for each SNP, which is the sum of the normalized X ("A" allele, Cy5 red) and Y ("B" allele, Cy3 green) intensities. We also used the B allele frequency (BAF), which is a modified version of the allelic intensity ratio theta ($\theta = 2/p*\arctan(Y/X)$), to reduce SNP-to-SNP variation in theta using the canonical genotype clusters.

I used Camin-Sokal maximum parsimony reconstruction implemented in PHYLIP [74] to infer the phylogenetic tree of SGA evolution at loci that showed variation in copy number or LOH between samples (informative loci).

2.3.2. Results and Discussion

DNA extracted from leukocytes or from a gastric sample represents an unaltered state of the genome

To detect SGA, a sample representing the normal, unaltered state of the somatic genome is needed. During an endoscopy, typically a blood sample and a biopsy from the upper part of the stomach are collected. The gastric biopsy is taken at around ~43-50cm endoscopic depth from the incisors of the patient, past the gastroesophageal junction (GEJ). I compared all four samples, samples #13, #14, #15, and #16, against each other and found concordance between the signal intensity genome-wide, and no apparent somatic alterations by taking log ratios between any two samples (data not shown). Ultimately, blood samples are comprised of red and white blood cells, where only white blood cells (leukocytes) contain DNA. And, DNA from leukocytes and DNA from gastric epithelium showed no somatic alterations suggesting that they can be used as control samples when evaluating DNA from epithelium isolated from biopsies sampled from the Barrett's segment.

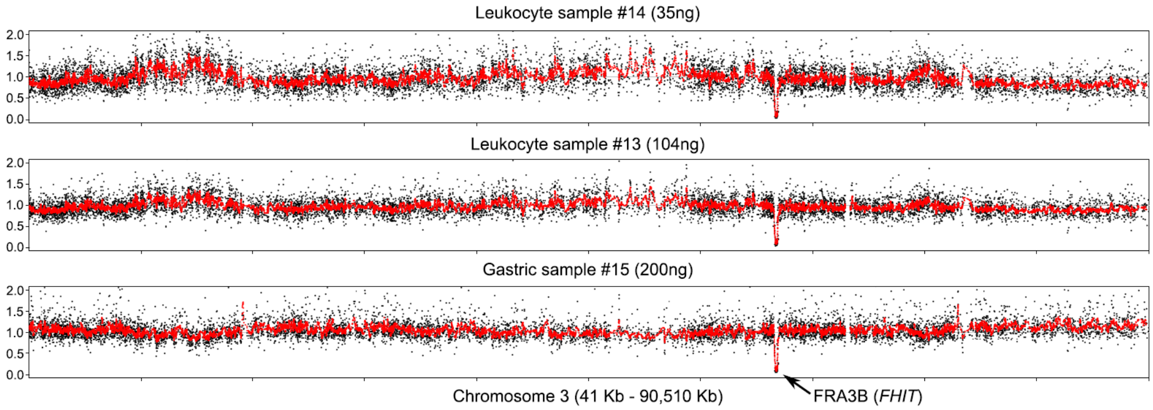


Figure 2.1. Ratio between R of two blood samples and one gastric sample used as reference samples and the same Barrett's sample. The signal profiles look similar and fragile site FRA3B is deleted in the Barrett's sample and intact in all 3 reference samples. The only difference between the reference samples is DNA concentration, where samples "blood 1", "blood 2", and "gastric" had 35ng, 104ng, and 200ng of DNA extracted and analyzed with Illumina 317K SNP arrays. The variation in the amount of DNA used for the genotyping assay induces a genomic waviness artifact similar to that described in Diskin et al. [75].

I showed that when taking the log ratio between the same BE biopsy and three different control samples, the same region of copy loss is detected, indicated by a negative log ratio. If any of the control samples were to have an alteration at the same location, a ratio close to zero would have been observed instead. Despite variability in signal intensity due to amount of input DNA was noted, this analysis determined that blood or gastric samples can be used as normal constitutive genotype control.

Correlation among GC-content of SNP probes, total signal intensity R from Illumina SNP arrays, and input DNA amount

As part of initial data quality assessment, I asked whether the %GC-content of SNP probes (the proportion of the bases G and C in the ~50-mer probe sequence) would affect the amount of signal intensity they produce. My expectation was that high GC content of a probe would allow better hybridization between the SNP probe and the sample DNA target due to three hydrogen bonds between GC as opposed to two between AT. As expected, I observed a strong correlation between probe GC-content and total normalized signal intensity R, which is the sum of the red and the green channel for both alleles of a given SNP (Figure 2.2). Luckily, the actual sequences of SNP probes were not proprietary and Illumina provided them free of charge, in a manifest file, which allowed this analysis.

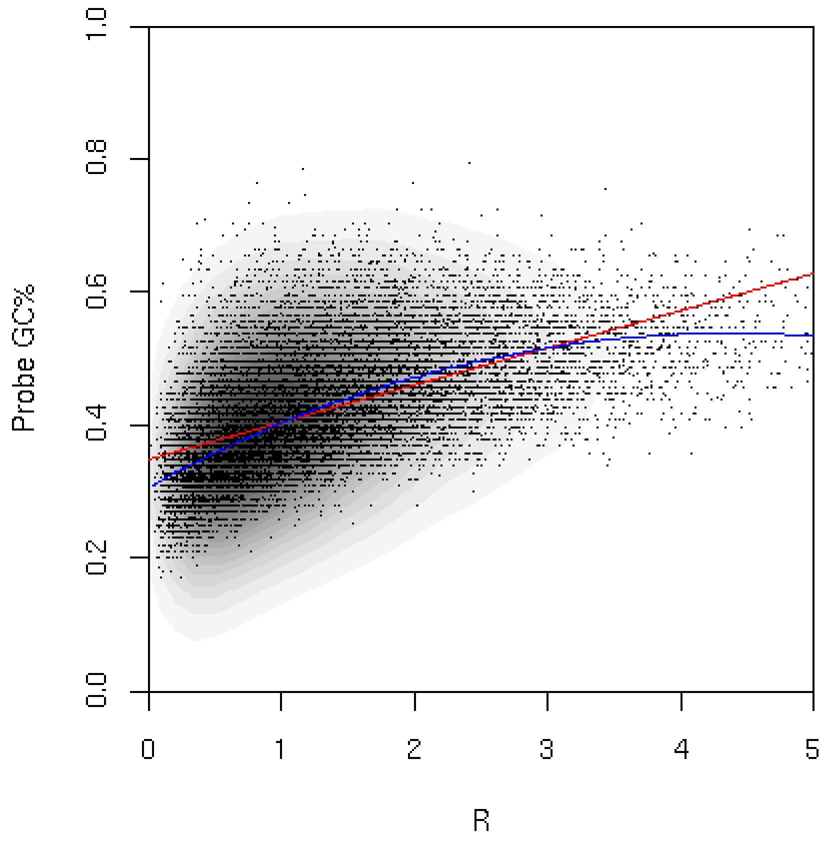


Figure 2.2. Total normalized signal intensity R increases if the SNP probes have higher GC-content. Shown are all SNPs on chromosome 1 from BE sample #1. The Pearson correlation between probe %GC and R was 0.57 for these data.

Having tested various input amounts of DNA, I observed that the correlation between probe %GC-content and total normalized signal intensity R decreased as input DNA amount increased (Figure 2.3). This suggested that as long as enough DNA is used to hybridize to the probes, the effect of how many Gs or Cs a probe contains to give a high signal diminishes. Illumina recommended using 750 ng of DNA for all analyses that use 317K SNP arrays; however, we aimed to minimize the amount of DNA we use per array as long as it gives enough SGA information in order to save as much biopsy material for further studies, as technologies advance. These analyses were performed in May 2008 and in Chapter 4 we used 200ng consistently across all arrays to minimize these effects, although

200ng would still potentially induce a 0.5 Pearson correlation between probe GC content and final total normalized signal intensity R.

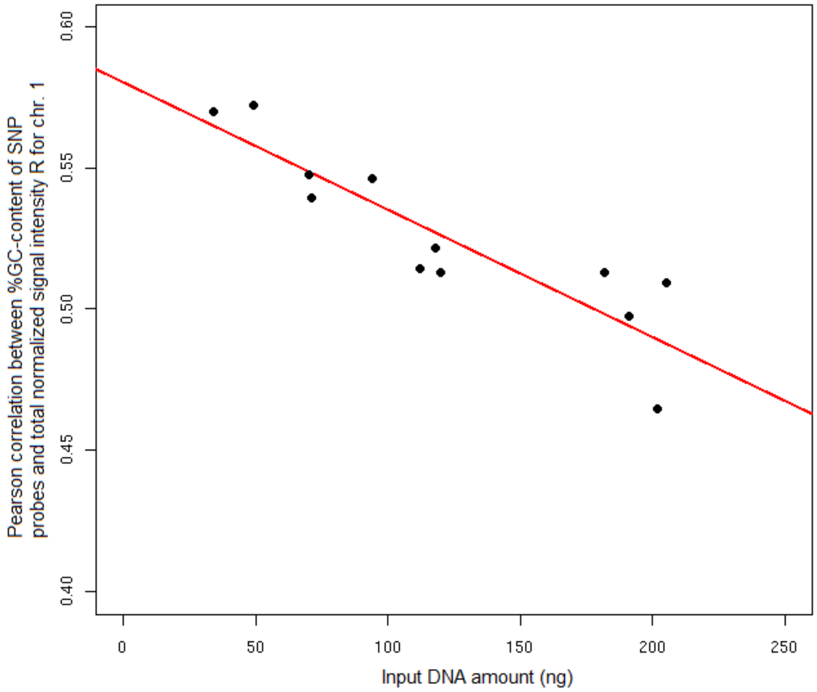


Figure 2.3. As input DNA amount increases, the correlation between probe GC content and total signal intensity R decreases. All 12 BE samples had various amounts of input DNA and at low amounts of input DNA (50ng or below) the signal from the arrays appears noisier and the effect of probe GC content was stronger. The Pearson correlation was -0.89 for these data. Note that the Pearson correlation for sample #1 that had 49ng of input DNA is 0.57 as shown in Figure 2.2.

Analyses of the DNA of samples #1-#12 revealed acquisition of somatic genomic abnormalities, where the most progressive copy number loss over time occurred at fragile site *FRA3B* that contains the gene *FHIT* (Figure 2.4,

Figure 2.5, and Figure 2.6) and at the genomic location of the tumor suppressor gene *CDKN2A* (Figure 2.6, Figure 2.7).

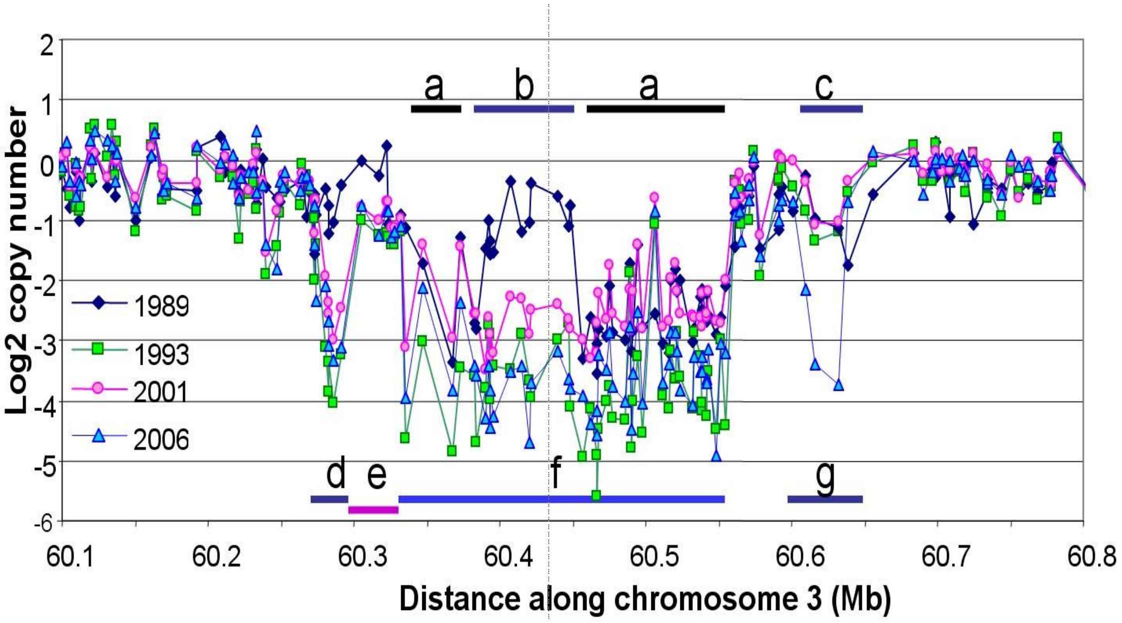


Figure 2.4. HumanHap300 (317k) array analysis of BE biopsies at the same level (±1cm) from one patient over 4 endoscopies from 1989 to 2006. In 1989 the BE consensus FRA3B region showed a region of 1-copy loss (b) flanked by 2-copy loss (a), with an adjacent region of 1-copy loss (c). In 1993 (a and b) merged into a region of uniform 2-copy loss (f), flanked by new regions of 2-copy (d) and 1-copy (e) loss. In 2006 the region of 1-copy loss at (c) lost its second copy (g). Adapted from Lai, Kostadinov, et al. [76].

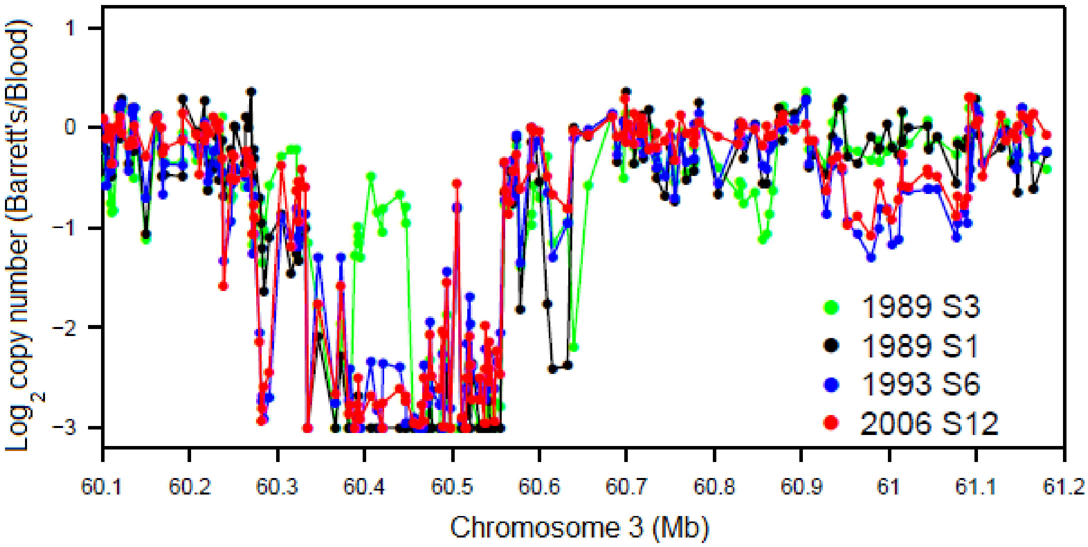
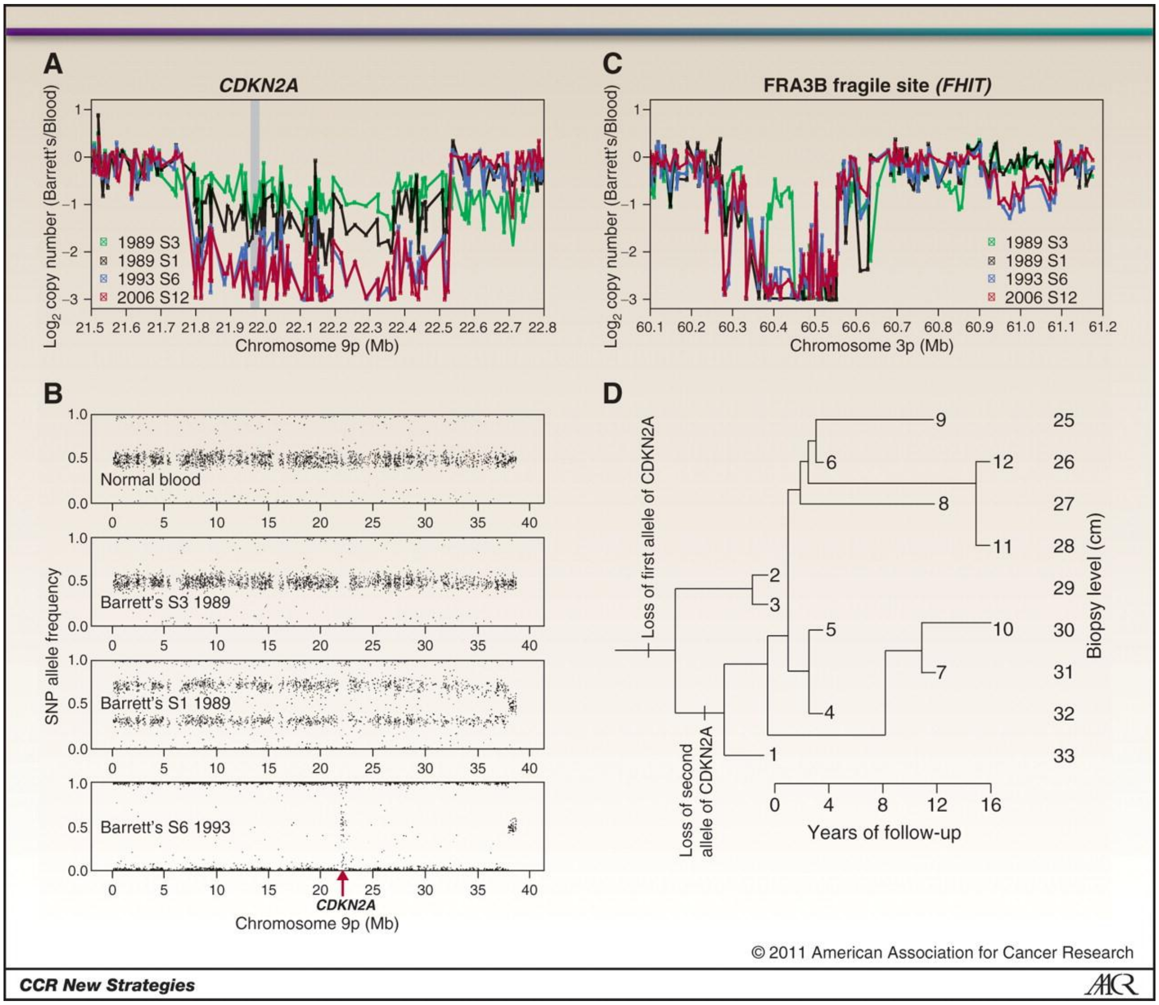


Figure 2.5. The same genomic region extended to 61.2 Mb, with a different set of samples from the same individual show progressive copy loss at various locations within the site. Adapted from (Brian J Reid, Kostadinov, and Maley 2011).



© 2011 American Association for Cancer Research

Figure 2.6. Benign clonal evolution in 1 patient with Barrett's esophagus studied longitudinally over 16 years. Purified Barrett's epithelium from endoscopic biopsies was assayed with Illumina 317K SNP arrays and compared with a blood sample control. A, Copy number analysis, normalized by SNP intensities from blood, reveals a single copy loss at CDKN2A in samples 2 (data not shown) and 3 in 1989, but homozygous deletion in CDKN2A in sample 1 and all samples from subsequent years. At first endoscopy in 1989, 2 clones were detected (1 with a small deletion of 1 allele at the CDKN2A locus, and the other with copy neutral LOH of the entire 9p arm with the CDKN2A deleted allele, generating biallelic deletion at CDKN2A). B, the SNP allele frequencies reveal a focal deletion in the CDKN2A locus in samples 2 and 3 in 1989, but sample 1 included a mixture of the clone from samples 2 and 3 with a new clone with copy neutral LOH of 9p and biallelic deletion of CDKN2A. All samples from 1993 and later show that the clone with biallelic deletion of CDKN2A went to fixation, leading to random noise in the allele frequencies for the SNPs in that region, as seen in the vertical ("waterfall") band in the bottom panel of B. The fact that the rest of the 9p arm remains diploid can be seen in the copy number data (A). C, The clone with deletion of the single allele of CDKN2A, which extends past 22.5 Mb on chromosome 9p, also had a single deletion in fragile site FRA3B at 60.42 Mb that

distinguishes it from the other clones. This and other lesions of the clone in samples 2 and 3 were not observed again after 1989, suggesting that this clone was driven to extinction by the clone from sample 1, with biallelic deletion of CDKN2A. D, A Camin-Sokal maximum parsimony reconstruction of the genealogy of clones based on the polymorphic copy number of lesions in 283 loci across the entire genome in the Barrett's biopsies shows that only one large clonal expansion occurred between 1989 and 1993. After 1993, the Barrett's segment remained stable, with accumulation of small interstitial lesions but no clonal expansions, no aneuploidy, and no progression to cancer. Figure adapted from Reid et al. [8].

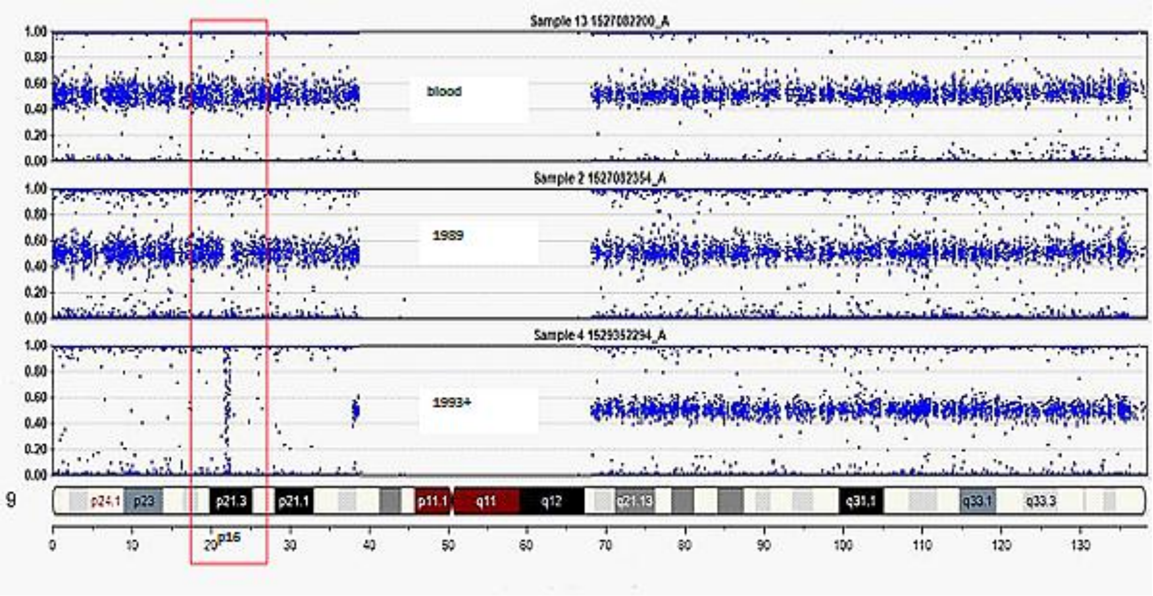


Figure 2.7. Output from Illumina's BeadStudio LOHPlus module showing allele frequencies in 3 samples. A normal sample (top) shows no LOH at p16 (red rectangle). In 1989 a small deletion of a single allele at p16 has appeared in all samples (middle). By 1993, all samples have lost both alleles of p16, with the original localized deletion (now apparent by a band of background readings with random allele frequencies) and loss of the entire arm of 9p in for other allele.

What is the best statistical measure of a SNP array technology's performance? The main purpose of SNP calling has been to associate genetic variants, or SNPs, with disease covariates, or in other words to detect which SNPs are at high linkage disequilibrium with an unknown disease marker. To this end, SNP arrays are mainly concerned with accurate calling of the genotype (the correct base pair) at a given SNP location, and the proportion of SNPs that have been assigned genotype calls (the "call rate") has been used as a metric for quality control and overall performance of the platform. However, for detecting copy number abnormalities, the call rate is not as relevant, since the primary aim is to detect regions of high or low probe fluorescence intensity in test samples in comparison to intensity of control samples. Therefore, instead of using call rates as metrics of platform performance, I focused on evaluating the capability of the platform to detect regions of DNA abnormalities and evaluating the effect of %GC content and input DNA amount on signal quality.

2.3.3. Conclusion

This single case of clonal evolution from longitudinal SGA data provided evidence that supports the following conclusions: 1) a second clonal expansion had occurred, associated with a double deletion in *CDKN2A*, after an initial single clonal expansion, associated with a single deletion in *CDKN2A* as well as copy number losses at *FHIT*; 2) for over ~12 years since the second clonal expansion, we did not observe a gradual accumulation of appreciable number of SGAs, except for small scale progressive losses at fragile sites, such as *FRA3B*; 3) DNA from leukocyte and gastric samples both showed no SGAs thereby proving being both useful as paired controls when evaluating SGAs in DNA of BE biopsies; and 4) using the similar amounts of input amount of DNA on a SNP array can minimize genomic waviness artifacts.

2.4. Cross-sectional meta-analysis of copy loss and loss of heterozygosity in Barrett's esophagus

Authors: Rumen Kostadinov, Xiaohong Li, Thomas G. Paulson, Patricia C. Galipeau, Brian J. Reid, Carlo C. Maley

2.4.1. Introduction

Barrett's Esophagus (BE) is a pre-malignant neoplasm that increases the risk of developing esophageal adenocarcinoma (EA) [77]. To identify groups of BE patients at high risk of cancer progression, we sought to identify common chromosomal aberrations across the full risk spectrum of the condition. I implemented a meta-analysis of three studies from the Seattle Barrett's Esophagus Project. The goal of this analysis was to combine SNP and array-CGH datasets of chromosomal loss from BE and EA samples to pinpoint regions of common loss across patients.

2.4.2. Methods

The three datasets included Illumina 33k SNP arrays on whole biopsies (34 patients) and surgical resections specimens (8 patients) from Li et al. [78], an Illumina 317K SNP array on 12 flow purified biopsies (1 patient) from [8] and a 4,500 spot bacterial artificial chromosome (BAC) hybridization array on 157 flow purified samples (72 patients) from Paulson et al. [79]. When there were multiple samples from a patient, I included the union of all detected lesions across those samples but only counted a lesion once per patient for the purposes of analysis. All SNP arrays were run on both BE and normal (gastric or lymphocyte) samples from the same patients for comparison. All BAC arrays were run on BE samples and compared against a common reference sample.

Illumina's BeadStudio software was used to call genotypes and produce signal intensity data in $\log_2 (R_{sub}/R_{ref})$ format that represents the difference in copy number of BE versus normal samples, where we assume normal samples have no aberrations. I then processed the SNP data to call regions of copy number loss using GLAD [80] setting logR ratio thresholds of -0.2 for single and -1.5 for double copy loss. BAC data was processed by a wavelet method [81] to call copy loss, copy gain or no aberration for every BAC. BAC data was analyzed by Xiaohong Li and the SBEP team and the final copy alteration and LOH data were provided to me. Regions of copy number loss, for the combination of both SNP and BAC datasets, were analyzed using STAC [82] to identify statistically significant areas of loss across samples. The STAC analysis was performed at 0.5Mb resolution using 500 permutations.

2.4.3. Results and Discussion

The combined STAC analysis identified 78 regions that were significant at the 95% confidence level, after multiple testing correction, including some previously known losses at chr. 3: 59-61MB (FHIT, FRA3B), chr.16: 77-77.5Mb (WWOX, FRA16D), chr. 9p: 21-32Mb (p16/CDKN2A/INK4a), and some newly discovered losses at chr. X: 31.5-32Mb (DMD), chr. 22: 22.5-23Mb (SMARCB1, DERL3, SLC2A11, MIF, GSTT1, GSTT2, DDT, CABIN1, SUSD2, GGT5) and chr. 18: 57-57.5Mb (CDH20) (Table 2.5, Figure 2.8).

	Chromosome	Start (Mb)	Stop (Mb)	Copy Loss p-value	LOH p-value	Genes				
Copy loss	1	242	242.5	0.032	0.002	ZNF238	ZNF238			
Copy neutral LOH	1	186.5	187		0.018	none				
Copy neutral LOH	1	199	200		0.018	many				
Copy neutral LOH	1	242.5	243		0.002	C1orf100	ADSS	C1orf101	FAM152A	
Copy neutral LOH	2	198.5	199		0.034	PLCL1				
Copy loss	3	59.5	61	0.002	0.002	FHIT				
Copy neutral LOH	3	61	62		0.002	PTPRG				
Copy loss	4	92	92.5	0.002	0.002	none				
Copy neutral LOH	4	118.5	119		0.002	none				
Copy neutral LOH	5	84	85		0.028	none				
Copy neutral LOH	5	99.5	100		0.028	FAM174A				
Copy loss	9	8.5	10.5	0.002	0.002	none				
Copy loss	9	12	12.5	0.002	0.002	none				
Copy loss	9	19	19.5	0.002	0.002	many				
Copy loss	9	20.5	28	0.002	0.002	many				
Copy neutral LOH	9	0	5.5		0.002	many				
Copy neutral LOH	9	6.5	8.5		0.028	GLDC	JMJD2C	C9orf123	PTPRD	
Copy neutral LOH	9	10.5	12		0.002	none				
Copy neutral LOH	9	12.5	13		0.002	TYRP1	C9orf150			
Copy neutral LOH	9	13	19		0.002	many				
Copy neutral LOH	9	19.5	20.5		0.002	SLC24A2	MLLT3			
Copy neutral LOH	9	28	30.5		0.002	none				
Copy neutral LOH	9	31	36.5		0.002	many				
Copy loss	10	68	68.5	0.016		LRRTM3				
Copy neutral LOH	11	83	83.5		0.002	none				
Copy neutral LOH	11	127	127.5		0.038	none				
Copy neutral LOH	13	83.5	85.5		0.018	SLITRK6				
Copy neutral LOH	15	50	50.5		0.002	many				
Copy loss	16	77	77.5	0.002	0.002	none				
Copy neutral LOH	16	7	8		0.004	A2BP1	A2BP1			
Copy neutral LOH	16	8.5	9		0.02	many				
Copy neutral LOH	16	77.5	78.5		0.002	MAF	MAF			
Copy neutral LOH	17	0.5	1		0.028	many				
Copy neutral LOH	17	1.5	4.5		0.028	many				
Copy neutral LOH	17	5.5	6.5		0.01	none				
Copy neutral LOH	17	8.5	10.5		0.01	none				
Copy neutral LOH	17	11	12		0.028	FLJ45455	DNAH9	DNAH9	ZNF18	MAP2K4
Copy neutral LOH	17	13.5	14		0.028	CDRT15P	COX10			
Copy neutral LOH	17	16.5	17.5		0.028	many				
Copy loss	20	14.5	15	0.006		none				
Copy loss	X	31.5	32	0.006		DMD				
Copy loss	X	152.5	153	0.002		many				

Table 2.5. Summary of 78 genomic regions that show copy number loss and copy neutral LOH significantly in common across studies.

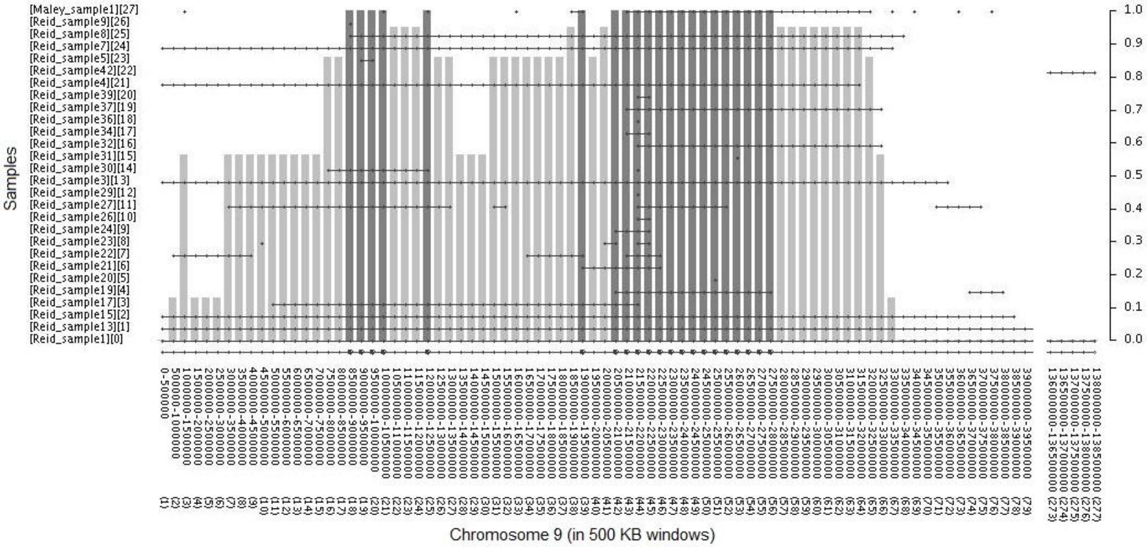


Figure 2.8. Example output from STAC from a subset of the 42 individuals that showed chromosome 9p alterations and analyzed with 33K-SNP arrays and the union of all alterations from 12 biopsies from one BE individual analyzed with 317K-SNP array. Gray

bars represent genomic regions on chromosome 9p that are significantly altered across samples. This plot excludes the BAC array data from 72 patients.

2.4.4. Conclusion

Combining copy number alteration and LOH data across studies in STAC increases sample size that increases power to detect statistically significant regions of copy number alteration and LOH across samples. We identified numerous commonly altered regions across individuals with Barrett's that could be further investigated for whether individual genes within those regions are implicated in promotion leading to clonal expansion or progression leading to esophageal adenocarcinoma.

2.5. Pilot experiment evaluating Affymetrix SNP6.0 and Illumina OmniQuad 1M SNP platforms

2.5.1. Introduction

The Illumina OmniQuad 1 Million SNPs array technology became available in late 2009 allowing evaluation of 1 million loci of the human genome from approximately 200ng of human DNA sample at a relatively low cost per array.

2.5.2. Methods

We (SBEP team members, CCM, RK) evaluated both Illumina OmniQuad 1Million-SNPs and Affymetrix SNP6.0 platforms in preliminary, pilot studies. About 42 DNA samples from epithelial-isolated BE biopsies were evaluated with Affymetrix SNP6.0 arrays that were run at the Vanderbilt Microarray Shared Resource (VMSR). A smaller set of DNA samples from epithelial-isolated BE biopsies were evaluated with Illumina OmniQuad 1M arrays at the Fred Hutchinson Genomics Facility.

During an endoscopy, the collected whole biopsies are frequently a mixture of epithelium and stroma, that is, a mixture of Barrett's epithelial cells, fibroblast cells, inflammatory cells, and occasionally normal squamous epithelial cells. The easiest assay to translate to the clinic is an assay that requires the minimum number of steps from biopsy collection to a readout of the somatic genomic abnormalities in the biopsy. Therefore, evaluating DNA from whole biopsies would be easier to translate to the clinic than evaluating DNA from epithelial-isolated biopsies. However, only the Barrett's epithelium typically contains somatic alterations and other cells present in the biopsy typically have a normal genotype and evaluating a whole biopsy mixture can reduce detection of SGA in BE cells. We designed a pilot experiment to compare the SGA detection between whole biopsies and epithelial-isolated biopsies. We adapted epithelial isolation method and technique from [73] and detailed protocol is given in the Methods section of Chapter 5.

SBEP collaborators processed biopsies from 8 BE individuals such that individual biopsies were split in half and DNA was isolated from whole half-biopsies and from epithelial-isolated half-biopsies. Also, DNA from paired blood and gastric samples was also evaluated for each BE individual to serve as constitutive genotype control. All biopsies were evaluated with Illumina OmniQuad 1M SNP platform. I compared signal quality and detection of SGA between 13 "epithelial-isolated versus whole" biopsy pairs.

During this pilot experiment, I developed a MySQL database to store all SNP data and associated sample information, and developed a pipeline of analysis using the R statistical language [83] and a browser-based visualization application using Linux/Apache/MySQL/Perl/PHP/Ajax and Perl-based Circos program by [84]. This set of software tools facilitated the development of SGA calling algorithms as well as storage and visualization of raw and processed SNP data.

2.5.3. Results and discussion

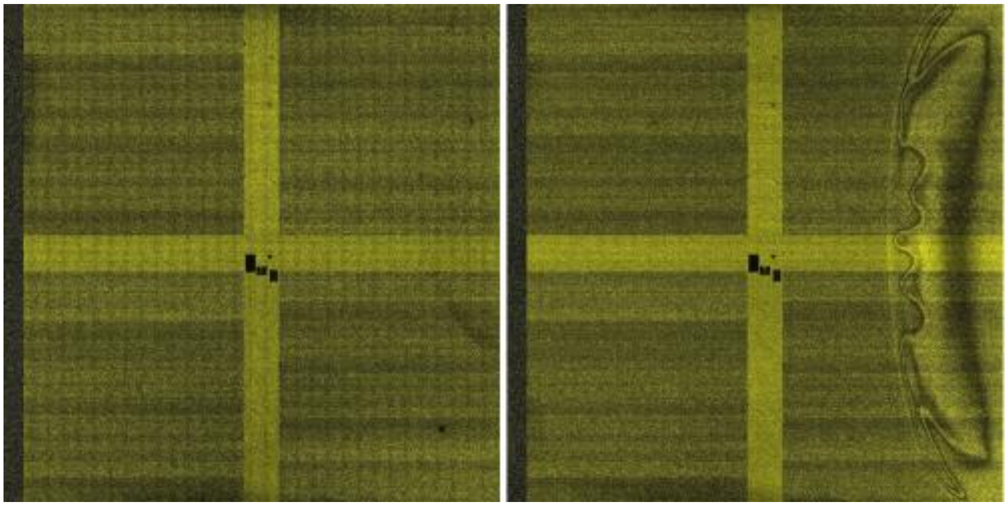


Figure 2.9. dChip-generated raw signal intensity images from two BE samples evaluated with Affymetrix SNP6.0 at the VMSR facility.

The cross separating the image into four quadrants is composed of copy number variation (CNV) probes and each quadrant is composed of SNP probes. Every quadrant shows consistent horizontal light and dark banding patterns. The first image also shows vertical dark stripes that run through the entire height of the image and that are approximately at even intervals horizontally. The second image shows a spatial artifact spanning the upper left and lower left quadrants that could be a manufacturing defect. The first image shows one dark scratch and one dark point in the lower right quadrant that are also spatial artifacts.

I observed horizontal light or dark banding patterns that were consistent across arrays (Figure 2.9). Occasionally, there were also horizontal dark lines crossing the extent of the array vertically (Figure 2.9). I suggested that this consistent variability is due somehow to the manufacturing process of the arrays, since Affymetrix arrays are printed using a sequence of masks, and each SNP probes would always have a predetermined spatial location on the array. VMSR collaborators suggested that this observed phenomenon would not affect SNP call rates and that probe redundancy and bioinformatic normalization methods would take care of the vast majority of the probe intensity variability. However, to the best of my knowledge, I could find only one study by Wan et al. that addressed these horizontal banding patterns and designed a normalization method to account for them [85]. Also, SNP array assays are primarily used for genome-wide association studies, where the call rate (the proportion of SNP probes on the array that give out a reliable genotype call) is a useful quality control metric. However, for calling SGAs, call rate is useful only for calling loss of heterozygosity accurately and the primary quality control metric for calling SGAs is accurate measurement of total signal intensity across probes.

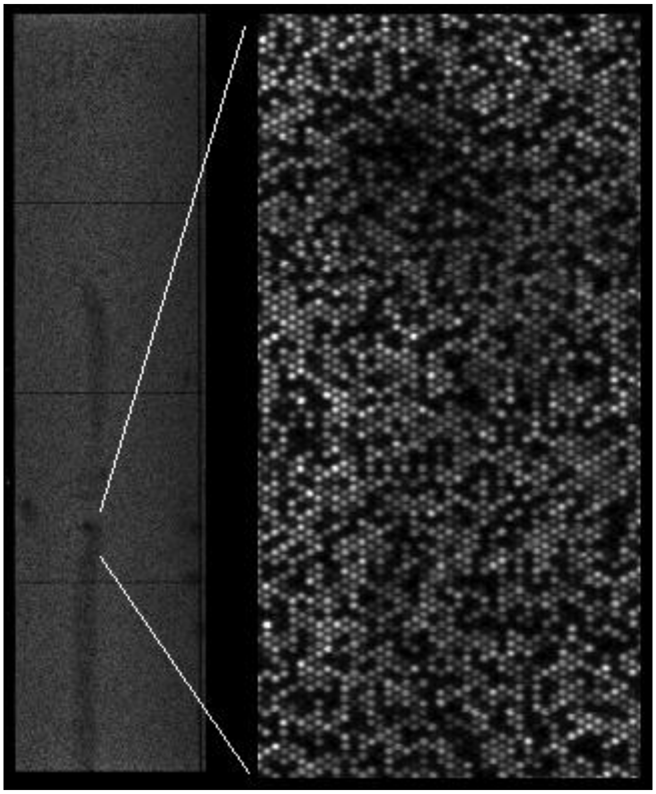


Figure 2.10. dChip-generated raw signal intensity image (from the green fluorescence channel) from four BE samples evaluated with Illumina OmniQuad 1M at the FHCRC genomic facility.

The image to the left is the actual 100% image that shows the hexagonal nature of the array.

The Illumina arrays also show some spatial artifacts, however due to the random placement of probes into hexagons on the chip during manufacturing [71], the signal from individual SNP probes would be random, as opposed to being consitently low or high corresponding to dark and light horizontal banding patterns from Affy SNP6.

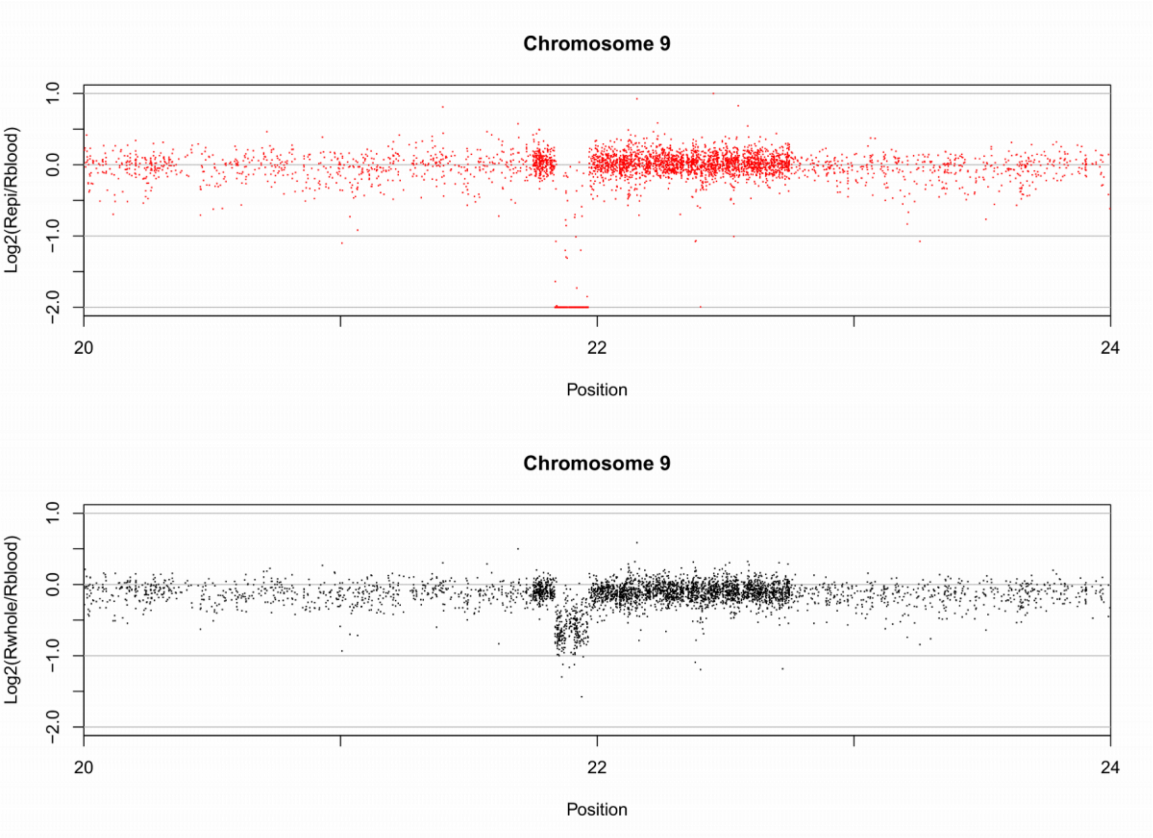


Figure 2.11. Double copy loss can be detected in an epithelial-isolated half-biopsy and can be miscalled as single copy loss in its paired whole half-biopsy. Every point represents a single SNP probe from 20-24 Mb on chromosome 9. On chromosome 9, the tumor suppressor gene *CDKN2A* lies between 21.96 and 21.99 Mb and both the epithelial-isolated half-biopsy (upper panel) and the whole half-biopsy (lower panel) show somatic loss in that chromosomal region. The log ratio between the epithelial-isolated sample and the blood control sample was -2.0 or lower, whereas the log ratio between the whole biopsy and the blood control sample was about -0.5. A two-fold lower signal intensity (log ratio of -2.0 or lower) in the epithelial-isolated sample would easily be called a double copy (homozygous) loss by an SGA calling algorithm, however a log ratio of -0.5 in the whole sample would be easily miscalled as a single copy loss.

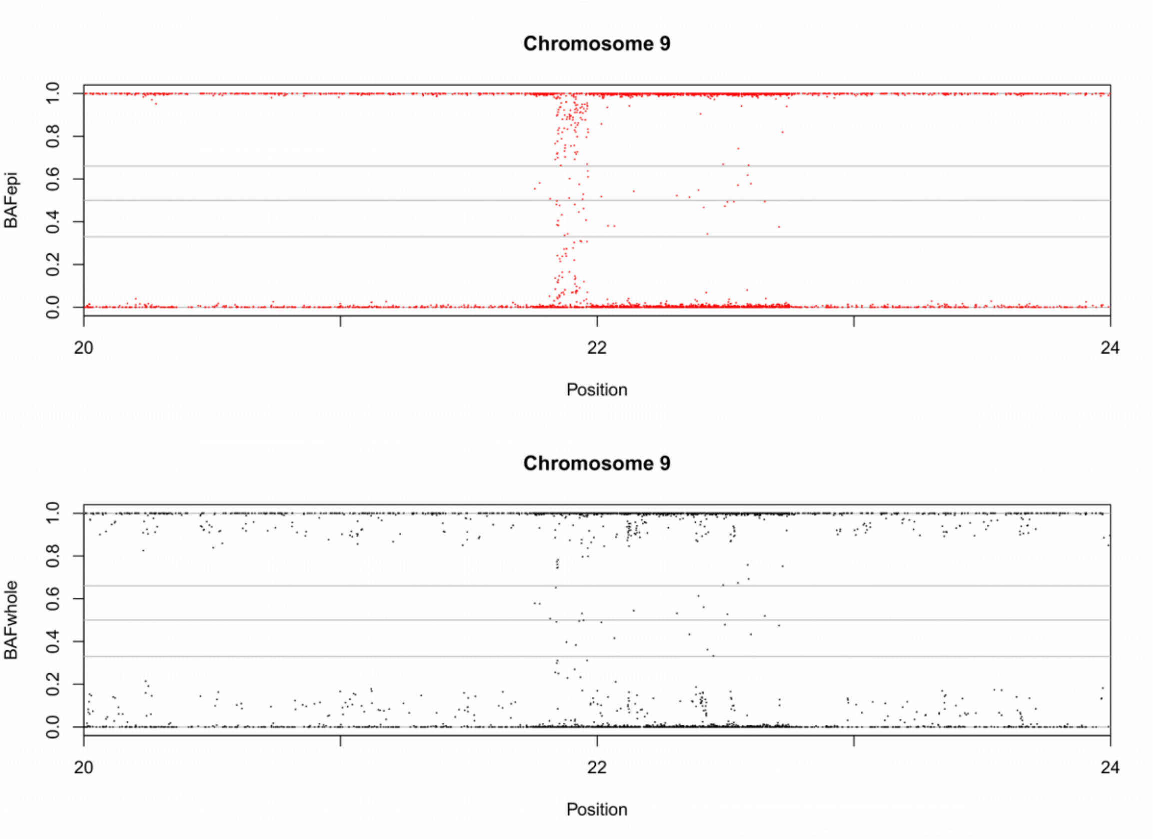


Figure 2.12. Copy neutral LOH can be detected in both epithelial-isolated half-biopsy and its paired whole half-biopsy. The B allele frequency of both half-biopsies show that the entire region 20-24 Mb of chromosome 9 has LOH, and since the signal intensity log ratio in

Figure 2.11 hovers around 0, this is indicative of copy neutral LOH. Notably, the epithelial-isolated half-biopsy shows a "waterfall" right in the *CDKN2A* region indicative that the signal from SNP probes within that region is so low that the scaled difference between the A allele intensity and the B allele intensity randomly fluctuates between 0 and 1.0.

Evaluation of whole versus epithelial-isolated biopsies using Illumina OmniQuad 1M

showed that probe intensity from whole biopsies was affected because of the mixture with

normal genotype cells and as result some SGAs would be missed (

Figure 2.11 and Figure 2.12). The detection of a "waterfall" pattern in B allele frequency

(Figure 2.12) suggests that the BE epithelial-isolated sample is primarily composed of BE epithelial cells and lacking any contaminating normal cells. If there was a significant contamination from cells or DNA having normal constitutive genotype a three cluster pattern, representing the AA, BB, and AB normal genotypes, would be observed instead of a "waterfall" pattern. This three cluster pattern is faintly visible in the whole half-biopsy in the *CDKN2A* region on chromosome 9 (lower panel of Figure 2.12).

2.5.4. Conclusion

Overall, same samples run on both platforms showed concordance in final SGA calls, which is also supported by a study by Curtis et.al. [86]. However, my assessment suggests that Affy arrays may suffer from consistent spatial biases in signal intensity requiring additional bioinformatic normalization steps, whereas Illumina may provide better initial signal intensity potentially requiring fewer bioinformatic normalization steps. Also, we had existing pipelines and expertise in handling and processing data with custom algorithms and software from the Illumina platform, which also influenced the choice of platform. The final choice of platform and facility was Illumina and FHCRC genomic facility, which had the advantage of having comparable cost for sample processing and the advantage of closer collaboration on DNA quantitation issues. Consequently, Illumina OmniQuad 1M was used for the large-scale experiment in Chapter 4. The final choice of biopsy processing method was to use epithelial isolation since using whole biopsies attenuated signal and resulted in missing or incorrect SGA calls.

2.6. Chapter summary

These four pilot experiments using Illumina and Affymetrix SNP array platforms helped provide enough preliminary evidence for accurate evaluation of SGA in BE biopsies and preliminary evidence for having a robust set of software tools to process SNP data and visualize results. The pilot experiment with 317K arrays on a single individual showed an initial clonal expansion, and a second clonal expansion, but an apparent evolutionary stasis of SGA after the second clonal expansion over the last 12 of a total of 16 years of clonal evolution. The pilot experiments comparing Affymetrix and Illumina 1M-SNP arrays suggested that both platforms performed approximately equally. The experiments also suggested that using epithelial isolation yields better signal compared to using a whole biopsy, which contains stroma in addition to Barrett's epithelium. All of the results served as preliminary evidence in winning a grant from the American Cancer Society that allowed us to perform a larger study of clonal evolution in 13 individuals and 161 biopsies using the Illumina 1M-SNP platform (Chapter 4).

Chapter 3. Agent-based model of evolutionary dynamics in Barrett's Esophagus

Authors: Rumen L. Kostadinov[1], Mary K. Kuhner[2], Kathleen Sprouffske[3], Lauren Merlo[4], Carlo C. Maley[5]

Author affiliations: [1]Genomics and Computational Biology Graduate Program, University of Pennsylvania; Department of [2]Genome Sciences, University of Washington; [3]Institute of Evolutionary Biology and Evolutionary Studies, University of Zurich; [4]Lankenau Institute for Medical Research; [5]Center for Evolution and Cancer, Helen Diller Family Comprehensive Cancer Center, Department of Surgery, University of California San Francisco;

Acknowledgements: I thank Dr. Todd Parsons for suggesting using a Gillespie algorithm for efficient simulation.

Author contributions: RLK, MKK, KS, LM, and CCM contributed to conception and design of the study. RLK, MKK, KS, and CCM contributed to design of all computational methods. RLK implemented all computational methods and generated results and plots of the data. RLK, MKK, KS, LM, and CCM contributed to analysis and interpretation of the data. RLK drafted the manuscript. RLK, MKK, KS, LM and CCM critically revised the manuscript.

Financial disclosure: Supported by NIH P01 CA91955, RC1 CA146973, R01 CA140657, R03 CA137811 and Research Scholar Grant #117209-RSG-09-163-01-CNE from the American Cancer Society

3.1. Abstract

Background

Esophageal adenocarcinoma (EA) is a disease of the somatic genome, the etiology of which relies on evolution by natural selection of somatic cells. In EA's precursor condition Barrett's esophagus (BE), increased genetic diversity at baseline detection is associated with increased risk of progression to EA. Clonal expansions have been observed in BE and have been thought to play a role in neoplastic progression. However, little is known whether and how clonal expansions affect genetic diversity dynamics, where dynamics is the change in the level of genetic diversity in a Barrett's cell population over time. Specifically, under what conditions can clonal expansions induce decreases in genetic diversity over time?

Methods & Findings

We assessed plausible scenarios for genetic diversity dynamics in BE with forward-in-time simulations of the change in genetic constitution of BE crypt populations over time. Initially, a BE segment was initialized with a genetically homogeneous population of crypts that fill out a hexagonal 2D lattice. During each simulation crypts mutate, replicate, die, and expand in neighboring hexes according to mutation rate, replication rate, death rate, and neighbor interaction parameters, respectively, based on their initially normal but subsequently mutated genotypes. Additionally, a duration of 20 years from initiation to EA and five rate-limiting mutations to EA were assumed per simulation run. We found that in less than 6% of parameter value combinations, genetic diversity decreased by more than 1% at any time during simulation runs. Decreases in genetic diversity are also magnified when we make the assumption that crypt mutation rate is independent of crypt replication rate as opposed to

proportional to it. We also found that neighbor interaction was one of the most important parameters for determining the speed of progression, i.e. whether or not a crypt must wait for a neighbor crypt to die before it can divide into the emptied space.

Conclusions

These modeling findings suggest that genetic diversity most likely increases monotonically over time despite clonal expansions. Therefore, we predict that monotonic increases in genetic diversity over time would be observed *in vivo* in most individuals with BE.

3.2. Introduction

Esophageal adenocarcinoma is a disease of the somatic genome, the etiology of which relies on evolution by natural selection of somatic cells. Although heterogeneity in premalignant conditions and cancer has been recognized, genetic diversity as measured in the fields of ecology and evolutionary biology has not been widely applied to cancer diagnosis and treatment. Little is known how genetic diversity within neoplasms changes over time, where competing hypotheses suggest that genetic diversity may either increase monotonically over time or decrease periodically during clonal expansions (see Figure 2 in [27]). Computational modeling can be used to understand genetic diversity as one fundamental aspect of progression to cancer in hope of advancing a conceptual understanding of progression, which may guide the design of future therapies to interrupt progression.

Although genetic heterogeneity is increasingly recognized as being a useful metric of neoplastic progression [31,87], to date, only a limited set of studies have highlighted an association of genetic diversity with clinical features of neoplasms. Barrett's esophagus is a unique pre-malignant human condition, in which the dynamics of genetic diversity can be studied over time *in vivo*. This is possible since current clinical practices recommend enrollment of individuals with BE in endoscopic surveillance programs, which collect and store biopsies from Barrett's tissue over time [5]. Maley et al. and Merlo et al. found that multiple measures of clonal genetic diversity in biopsies, collected at initial (baseline) detection of Barrett's, predict progression to EA [88,89]. Genetic diversity has also been measured in other cancer types. Park et al. characterized diversity in 15 human breast tumors, which were diverse mixtures of ductal carcinoma in situ and invasive regions, and found that genetic diversity, measured using Shannon's index, was associated with clinically relevant variables, such as tumor grade [33]. If genetic diversity in neoplasms is associated with clinically relevant variables, understanding how the level of genetic diversity changes over time may help in estimating the timing of manifestation of clinically relevant features, or in other words, in measuring the onset of clinical stages of carcinogenesis.

Clonal expansions characterize the growth of clonal cell populations, yet few studies have shown that measurements of clonal expansions can predict progression of a pre-malignant condition to cancer. Galipeau et al. observed clonal expansions in 61 individuals with BE by measuring loss of heterozygosity at microsatellite loci on chromosomes 9p and 17p, where some expansions had covered the entire BE segment [90]. Based on these data, Galipeau et al. hypothesized that clones bearing genomic abnormalities such as LOH can arise independently or bifurcate over time, which generates observed clonal heterogeneity within BE segments [90]. Leedham et al. found heterogeneity at individual crypt level where crypts within microdissected blocks showed similar LOH and point mutation patterns, and crypts from spatially distant blocks showed distinct LOH and point mutation patterns, suggesting local clonal expansions and a degree of clonal intermingling [34]. Graham et al.

hypothesized that the generation and maintenance of genetic diversity in Barrett's is explained by 1) strong mutation and strong selection, 2) group selection, where diverse clones as a group are selected, and 3) an hidden initial landscaping genetic abnormality triggering clonal expansion on which neutral mutations occur and generate diversity [57]. Clonal expansions have been measured in pre-malignant ulcerative colitis, which can progress to colorectal cancer. Salk et al. measured the lengths of polyguanine microsatellites in biopsies collected from 2D patches of pre-malignant tissue in patients with ulcerative colitis and found that clonal expansions inferred by shared microsatellite length patterns were associated with progression to colorectal cancer [51]. Salk et al. further hypothesized that clonal expansions measured with any neutral markers could be useful for predicting neoplastic progression [50].

The interplay of genetic diversity and clonal expansions is complex. Genetic diversity is an outcome of the underlying evolutionary parameters of neoplastic cell populations, including mutation, selection, clonal expansion, and spatial structure. Martens et al. explored the effect of spatial structure on the waiting time to cancer and found that under most biologically realistic parameter values clones tend to arise independently and come in contact as they expand, slowing dynamics and reducing the time to cancer, as opposed to arising sequentially after clonal fixation, which leads to quicker progression to cancer [56]. Mutation rates have only been indirectly estimated from cancer age-incidence data in a few cancers, such as colorectal cancer [91], (never in Barrett's esophagus) and the number of clonal expansions involved in neoplastic progression have been estimated from 2 in lung cancer [92] to 20 in colorectal cancer [93] (there are no formal estimates in Barrett's esophagus). Here we explore how mutation rate and spatial structure parameters affect clonal expansion and genetic diversity by assuming a fixed fitness landscape with 5 fitness-increasing rate-limiting mutations that underlie at least 5 clonal expansions.

Also, we model "selective" (driver) and "neutral" (passenger) genetic abnormalities that represent any category of mutation, such as point mutations, copy number alterations, loss of heterozygosity, and structural rearrangements. Abnormalities at selective loci result in advancing tumor progression through the stages of carcinogenesis, and have a fitness increasing effect. Abnormalities at neutral loci confer no fitness benefit; however it is neutral loci that define the observed mutation states when sampling tumors, or in other words, neutral loci are observed variables that are dependent on the hidden (latent) variables of the underlying evolutionary parameters, such as mutation rates and fitness landscape defined by the fitness-modulating selective loci.

What are the dynamics, or change in magnitude, of genetic diversity over time? Our prior hypothesis is that in the absence of clonal expansions genetic diversity is a monotonically increasing function of time as new somatic genomic abnormalities (SGA) accumulate in the neoplastic cell population. However, we aim to explore the plausibility of decreases in genetic diversity during periods of clonal expansions that may homogenize the genetic constitution of the population. The population genetic diversity may be a monotonically increasing function of time, which can be modeled with linear, exponential, logarithmic, and other mathematical growth functions. However, the population genetic diversity may instead oscillate over time, for instance, it may increase as SGA accumulate but decrease during periods of clonal expansions, making genetic diversity harder to model, as Merlo et al. proposed [27].

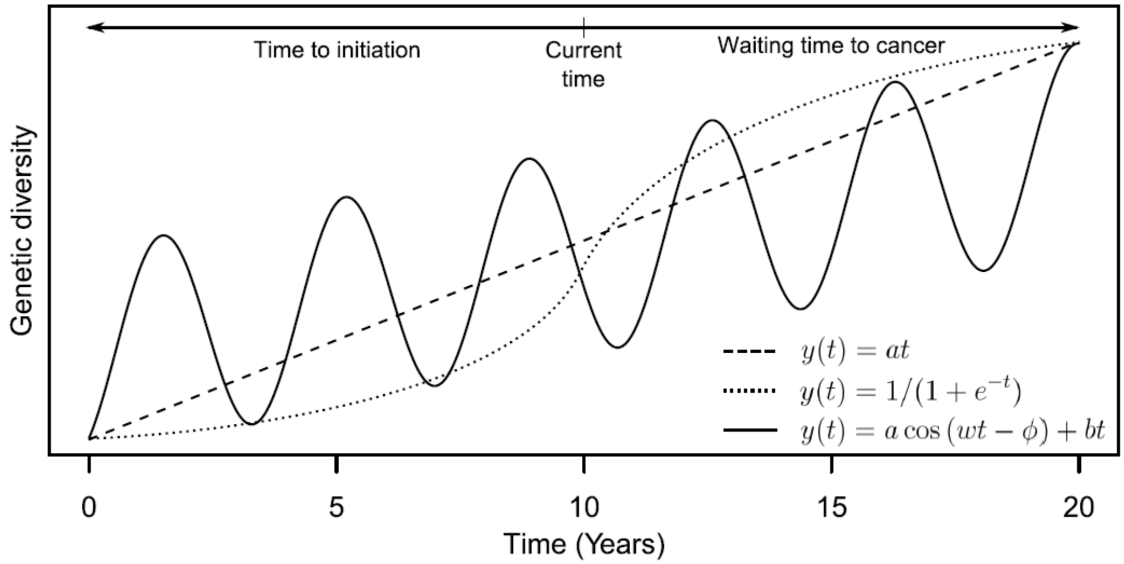


Figure 3.1. Plausible scenarios for genetic diversity dynamics. Genetic diversity can decrease periodically due to clonal expansions; for instance, if five necessary, sufficient, and rate-limiting steps (mutations) are acquired sequentially in the neoplastic cell population and boost fitness enough to cause five clonal expansions, they may homogenize the neoplasm and produce five troughs in genetic diversity (solid line). Such oscillating dynamics can be captured only with longitudinal data so that troughs in genetic diversity can be detected. Alternatively, genetic diversity can increase monotonically over time and modeled as a linear function of time (dashed line). Alternative hypotheses, such as a single genetic catastrophe or a critical phase transition point, can also be plausible, in this example, reduced to a simple sigmoid curve (dotted line). Importantly, at time of detection of a new tumor, measurement of genetic diversity and prior knowledge of the typical dynamics of genetic diversity taken together can estimate the time elapsed since tumor initiation and predict the waiting time for tumor progression to cancer. Elapsed time since initiation and waiting time to cancer are clinically relevant variables for diagnostic purposes and for weighing treatment options.

3.3. Methods

We fix a set of parameters θ at the beginning of each simulation $\theta = \{h, w, p, t, b, d, a, n, m, \mu, \nu, \mathbf{r}, \mathbf{s}, x\}$. The Barrett's segment is represented as a two dimensional $h \times w$ (height × width) hexagonal lattice, wrapped around along the h dimension to form a cylinder. At the beginning of the simulation, all $h \times w$ hexes are occupied by crypts having no mutations. Crypts have an initial birth rate b (number of crypt fission events per day), and when dividing, one of its daughter crypts remains in its original hex (self-renews) and the other daughter crypt either expands into an empty neighboring hex or, if none exist, kills off a neighboring crypt and occupies its hex space with probability p. Intuitively, p describes the stochastic neighbor contact process, where for $p > 0$, replicating crypts crowd out and replace their neighbor crypts, instead of waiting for an empty hex as when $p = 0$. Crypts have an initial death rate d (number of death events per day), where upon dying, the crypt's hex is emptied.

The mutation states of crypts are stored in matrix x, that has $h \times w$ rows (crypts) and $m + n$ columns, where m is the number of selective loci (modulating crypt reproduction and survival) and n is the number of neutral loci (having no effect on crypt reproduction or survival). Taking both neutral and selective mutations together, there are l = m+n total possible loci at which mutations may occur.

$$x_{i,*} = \{\underbrace{x_{i,1}, x_{i,2}, \ldots, x_{i,m}}_{m}, \underbrace{x_{i,m+1}, x_{i,m+2}, \ldots, x_{i,m+n}}_{n}\}$$

At the beginning of each simulation, all $x_{i,\ell} = 0$, and during the run, every mutation hitting a neutral locus $\ell \in \{(m+1), \ldots, (m+n)\}$ sets $x_{i,\ell} = x_{i,\ell} + 1$, to count multiple hits at neutral loci, whereas every mutation hitting a selective locus $\ell \in \{1..m\}$ sets $x_{i,\ell} = 1$, which makes selective loci irreversible, that is, when mutated they stay mutated for the duration of the run, whereas neutral loci can record the number of changes, which can be useful for representing microsatellite shifts, or single base pair substitutions.

Selective loci are stored in vectors $\mathbf{r} = r_1, \ldots, r_m$ and $\mathbf{s} = s_1, \ldots, s_m$, and each locus $\ell \in \{1..m\}$ must have either reproductive selection coefficient satisfying $r_\ell \in (-1, 0) \cup (0, \infty)$ or survival selection coefficient satisfying $s_\ell \in (-1, 0) \cup (0, \infty)$, otherwise $r_\ell = 0$ and $s_\ell = 0$ make the locus neutral. Thus, loci having $r > 0$ or $s > 0$ are selectively advantageous (increase birth rate and decrease death rate, respectively) and loci having $r < 0$ or $s < 0$ are selectively deleterious (decrease birth rate and increase death rate, respectively).

As crypts acquire mutations in selective and neutral loci over time, their birth, death, and mutation rates are calculated as follows:

$$\text{Equation 1:} \quad b_i = b\prod_{\ell=1}^{m}(1 + x_{i,\ell} r_\ell)$$

The birth rate b_i of crypt i is the initial crypt birth rate b multiplied by the reproductive advantage r_ℓ conferred from any mutated $x_{i,\ell} = 1$ selective locus ℓ.

$$\text{Equation 2:} \quad d_i = d\prod_{\ell=1}^{m}(1 - x_{i,\ell} s_\ell)$$

Similarly, the death rate d_i of crypt i is the initial rate of crypt death d multiplied by the survival advantage s_ℓ conferred from any mutated $x_{i,\ell} = 1$ selective locus ℓ. Note that a positive value for s_ℓ lowers the death rate, whereas a positive value for r_ℓ increases birth rate.

We allow neutral loci and selective loci to have distinct rates of mutation $\nu \neq \mu$, since, depending on the assay modeled, neutral loci n could represent somatic copy number polymorphisms, microsatellite or methylation sites, that have higher mutation rates $\nu \in [10^{-4}..10^{-6}]$. Other genetic alterations, such as point mutations and structural rearrangements could also be modeled with different mutation rates, but we have focused on alterations with relatively high mutation rates that are more likely to be used for quantifying within neoplasm diversity. However, those alterations are modeled in the selective loci m which represent the loci that will acquire the ~3-5 rate-limiting mutations necessary for initiation and promotion of epithelial cancers. Those rate-limiting steps (driver mutations) are thought to have relatively low mutation rates $\mu \in [10^{-6}..10^{-10}]$.

Equation 3: $$\mu_i(\tfrac{mutations}{day}) = m(\tfrac{loci}{1}) \cdot \mu(\tfrac{mutations}{locus \cdot birth}) \cdot b(\tfrac{births}{day}) \cdot \prod_{\ell=1}^{m}(1 + x_{i,\ell} r_\ell)^a$$

Equation 4: $$\nu_i = n\nu b \prod_{\ell=1}^{m}(1 + x_{i,\ell} r_\ell)^a$$

Equations 3 and 4 give the mutation rate at selective loci μ_i and neutral loci ν_i for crypt i. One of two mutation models is fixed at the beginning of each simulation: time-dependent $(a = 0)$ or replication-dependent $(a = 1)$. In the time-dependent model, the crypt mutation rate is not affected by changes in crypt division rate: setting the term $\prod_{\ell=1}^{m}(1 + x_{i,\ell} r_\ell)^0 = 1$ in Eqs. 3 and 4, whereas in the replication-dependent model, mutations modulating the crypt division rate, also modulate mutation rate using the term $\prod_{\ell=1}^{m}(1 + x_{i,\ell} r_\ell)^1$. Intuitively, the replication-dependent model means that crypts that divide faster will also mutate faster, reflecting more (cell-intrinsic error-prone DNA) replication of stem cells in dividing crypts, whereas the time-dependent model means that regardless of faster or slower crypt division, mutations are time homogeneous, as might occur with on-going stem cell replication in non-dividing crypts, as well as chronic DNA damage from acid reflux, inflammation, or other cell-extrinsic causes.

Thus far, equations 1-4 describe the four actions that a crypt executes with four distinct rates: division, death, mutation of a selective locus chosen at random, and mutation of a neutral locus chosen at random. Having a population of $h \times w$ crypts, each having four rates and distinct mutation states, presents a computational challenge for efficient forward-in-time stochastic simulation. We address that by using a Gillespie algorithm that can be summarized in three steps: choosing the first action to execute from an exponential distribution with rate equal to the sum of all action rates of the population of crypts, advancing time according to that action, and choosing the appropriate individual crypt that performs the action. When a crypt divides in space, its genotype must be copied to the new location, which is computationally intensive for a single simulation of ~0.1-1 billion crypt divisions. We address that by having hex locations point to crypts stored in a large phylogenetic tree in memory that records exact divergence times and mutation differences among clonal populations of crypts, resulting in average run times of ~5 minutes per run on Core i7-975 3.3 GHz, using ~3 GB of RAM.

Equation 5: $$\hat{\kappa} = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{\ell=m+1}^{m+n} |x_{i,\ell} - x_{j,\ell}|}{\binom{k}{2}}$$

We ran 360 simulations that represent all combinations of initial parameters in the fixed initial value column of Table 3.1. We also ran an additional 120 simulations (All parameter sweeps in Table 3.1, except that neutral mutation rate was fixed to 10^{-4}) to evaluate genetic diversity dynamics in a fully-mixed crypt population scenario, in which crypts divide into six randomly chosen hex locations in the grid according to the same rules as dividing into six neighbor hex locations: first, if any of the six locations are empty, divide into a randomly chosen empty hex, and second, if all of the six locations are occupied, replace a crypt with probability p. The state of the hexagonal 2D lattice was output at half-year intervals during each simulation run. The estimate of genetic diversity $\hat{\kappa}$ (kappa) was computed by sampling k

crypts (k = 100) from random hex locations without replacement from the $h \times w$ segment and summing the hamming distance between mutation states over all neutral loci $\ell \in \{(m+1), \ldots, (m+n)\}$ for all $\binom{k}{2}$ pairwise crypt comparisons (Equation 5). We ran a peak and trough detection algorithm to find local minima and maxima of the resulting curve of genetic diversity over time.

In this simulation, individual crypts occupy positions in a hexagonal grid. They have base rates of birth and death which can be modified by mutations at selected loci. Two key parameters of this system are whether crypts can kill their neighbors in order to reproduce or must wait for an empty space, and whether the rate of mutation is proportional to time or to crypts divisions (so that faster-dividing crypts, assumed to be driven by faster dividing stem cells, accumulate more mutations). Non-selected (neutral) loci are also modeled so that the simulation can be compared to genome-wide samples from Barrett's Esophagus.

Parameter	Description	Units	Fixed initial value	Does parameter value change during simulation run?	Range during simulation run
h	grid height	number of hex cells	300	no	N/A
w	grid width	number of hex cells	300	no	N/A
p	neighbor hex cell replacement probability	[0,1]	{0,.25,.5,.75,1}	no	N/A
t	simulation	days	0	yes	[0,7300]

	run time				
b	crypt birth rate	number of crypt fission events per day	0.02	yes	[0.005,1]
d	crypt death rate	number of crypt death events per day	0.016	yes	[0.005,0.1]
a	time-dependent mutation model (a=0) replication-dependent mutation model (a=1)	{0,1}	{0,1}	no	N/A
n	number of neutral loci	number of neutral loci	100	no	N/A
m	number of selective loci	number of selective loci	10	no	N/A
μ	crypt mutation rate of selective loci	number of mutations per locus per crypt birth	{10^{-6}, 10^{-7}, 10^{-8}}	no	N/A
v	crypt mutation rate of neutral loci	number of mutations per locus per crypt birth	{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}}	no	N/A
r	vector of reproduction selective advantage	{[0,∞),…, [0,∞)}	{.6,.8,1,1.2,1.4, 0,0,0,0,0}	no	N/A

	multipliers				
s	vector of survival selective advantage multipliers	{[0,∞),…, [0,∞)}	{0,0,0,0,0, .6,.8,1,1.2,1.4}	no	N/A
X	Matrix of mutation states at neutral and selective loci (columns) of all crypts (rows)	For each selective locus 1..m the state is {0,1} For each neutral locus m+1..m+n the state is [0..∞)	0	yes	**X**[*,1..m] ranges {0,1} **X**[*,m+1..m+n] ranges [0..∞)

Table 3.1. Parameter configurations for forward-in-time simulation of evolutionary dynamics in Barrett's Esophagus.

3.4. Results

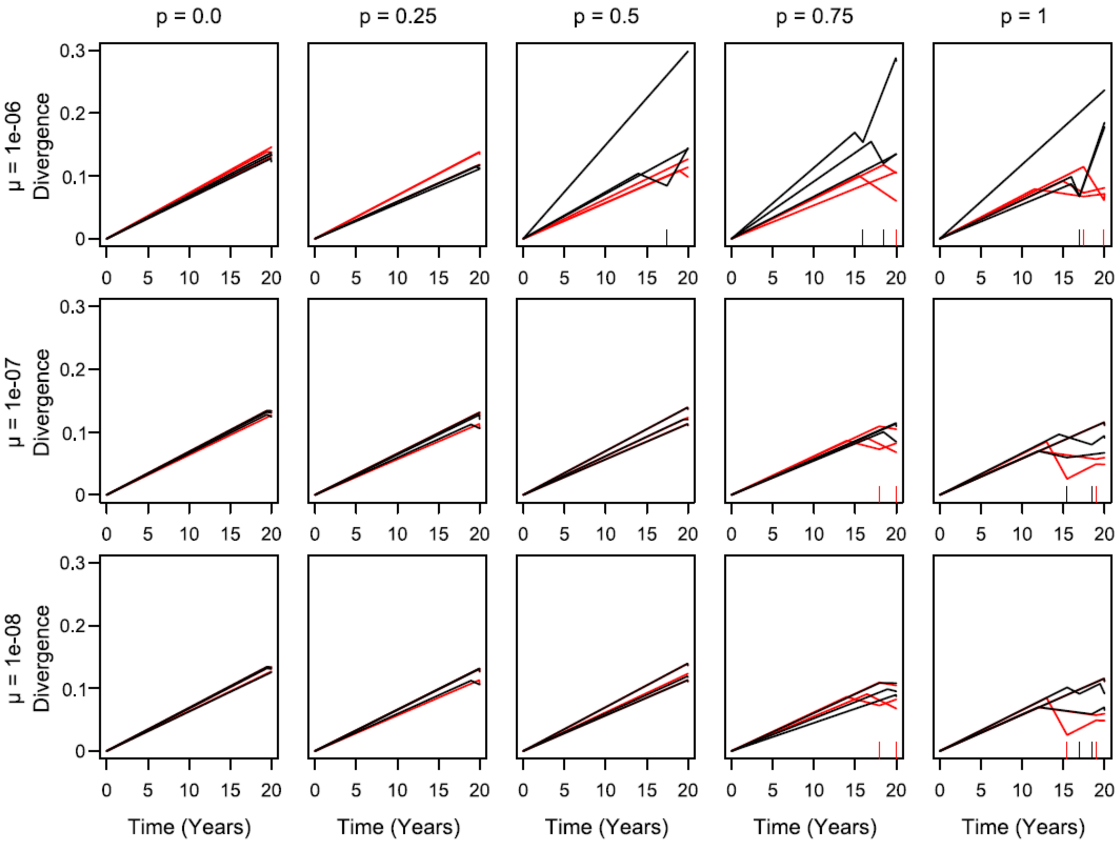


Figure 3.2. A spatially-structured crypt population induces monotonic increase in genetic diversity. Genetic diversity is computed by taking a random sample of 100 individual crypts from the grid. The neutral mutation for the displayed parameter sweeps is 10^{-4}. Black lines represent crypt division-dependent mutation (a=1) and red lines represent time-dependent mutation (a=0). Each parameter sweep was run in triplicate.

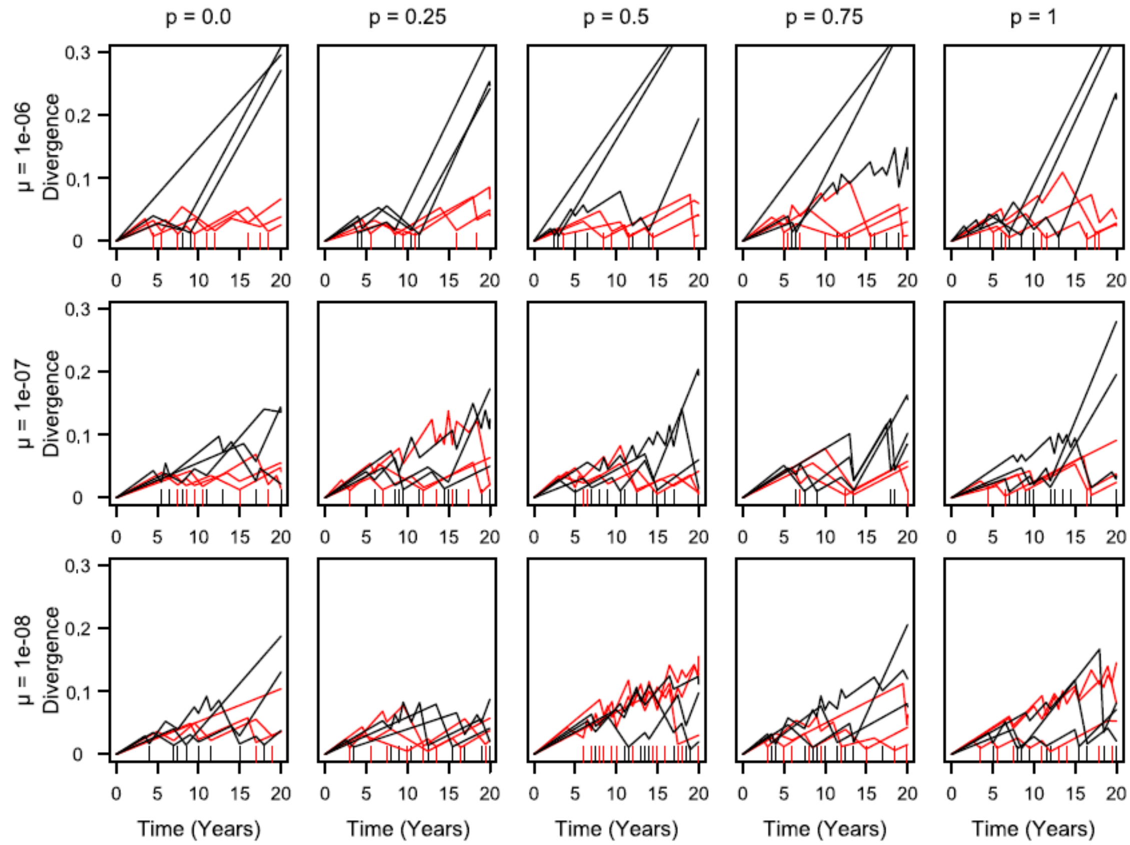


Figure 3.3. A fully-mixed crypt population induces periodic decreases in genetic diversity. Genetic diversity is computed by taking a random sample of 100 individual crypts from the grid. The neutral mutation for the displayed parameter sweeps is 10^{-4}. Black lines represent crypt division-dependent mutation (a=1) and red lines represent time-dependent mutation (a=0). Each parameter sweep was run in triplicate.

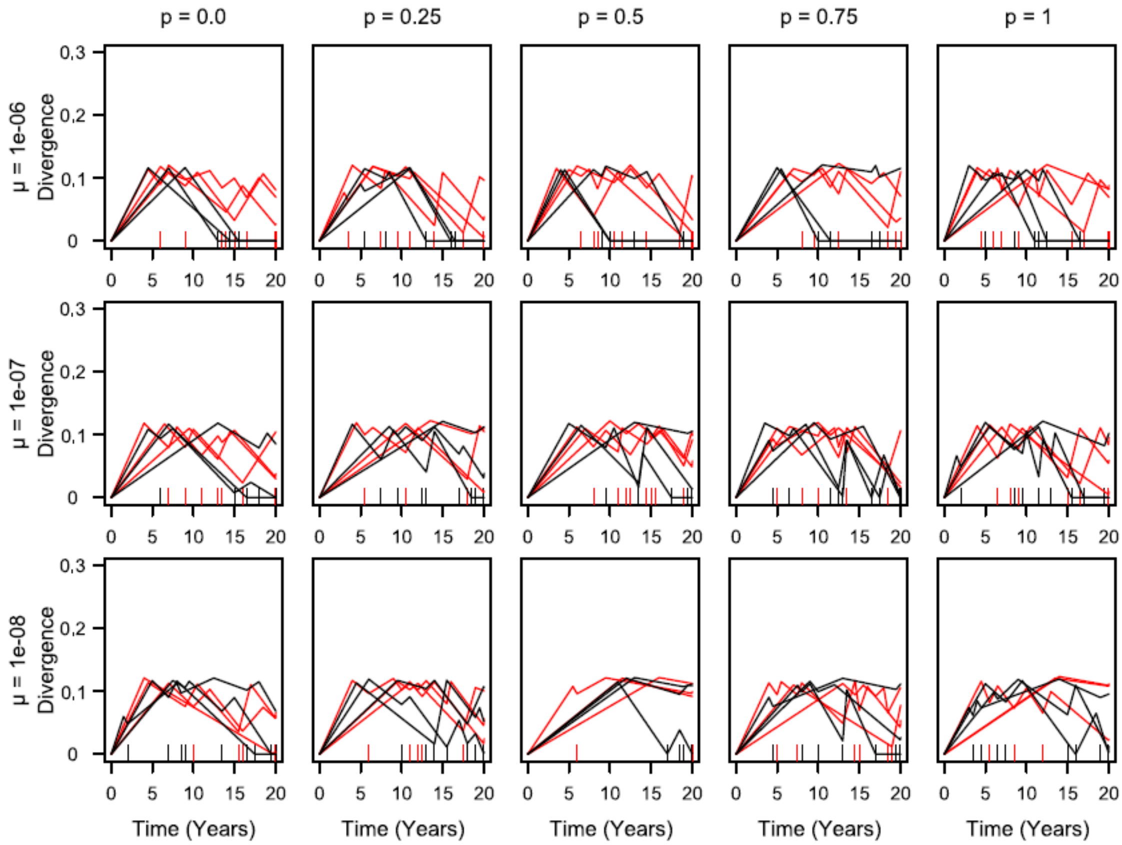


Figure 3.4. Biopsy sampling (n=12) for a fully-mixed crypt population, where biopsy size was 10x10 crypts and mutation detection threshold was 0%.

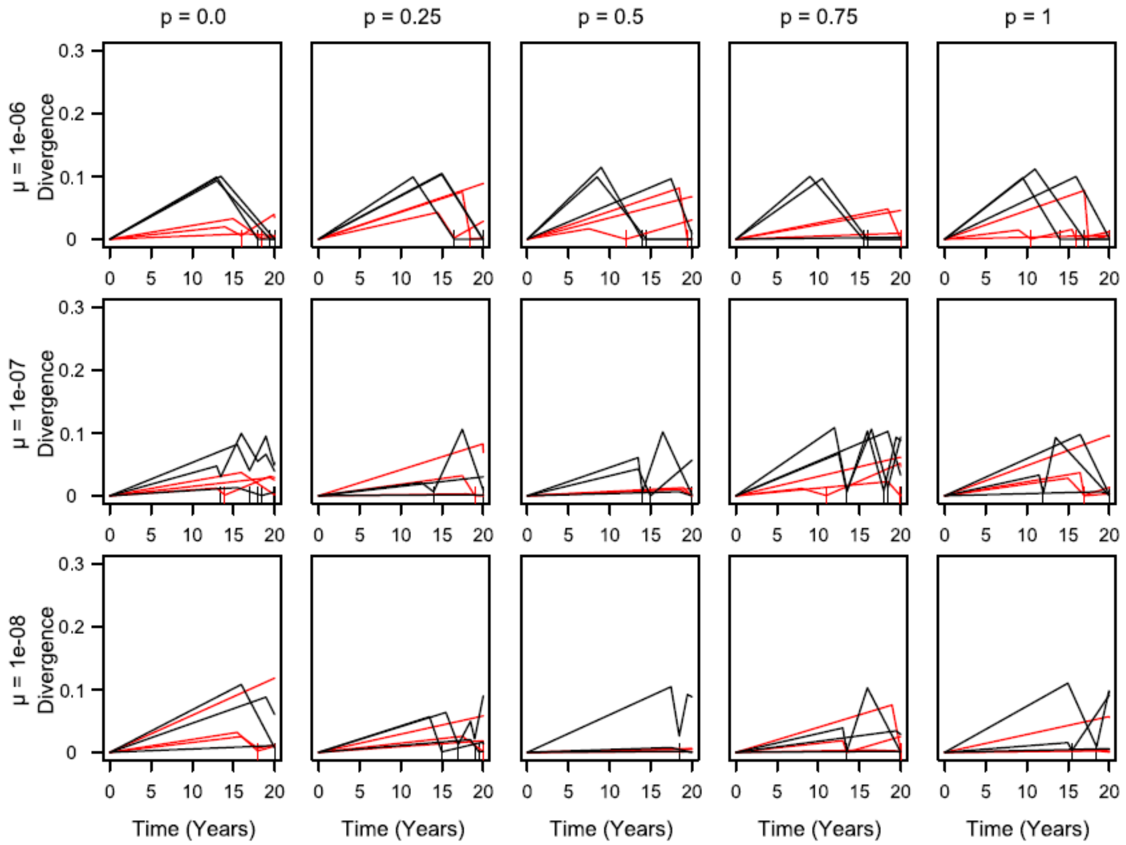


Figure 3.5. Biopsy sampling (n=12) for a fully-mixed crypt population, where biopsy size is 10x10 crypts and mutation detection threshold is 5%. The neutral mutation for the displayed parameter sweeps is 10^{-4}. Black lines represent crypt division-dependent mutation (a=1) and red lines represent time-dependent mutation (a=0). Each parameter sweep is run in triplicate.

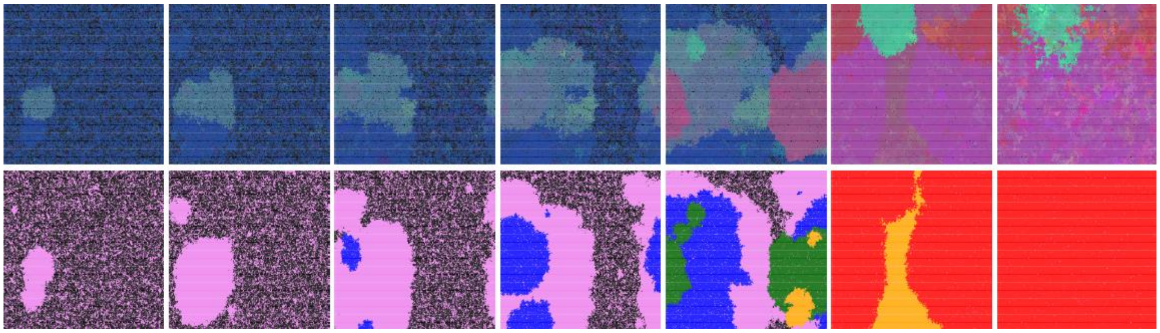


Figure 3.6. Example grid spatial dynamics from year 5-20 (7 panels left to right), when selective mutation rate is 10^{-6} and neutral mutation rate is 10^{-4}. A fully-mixed crypt population has quicker genetic diversity dynamics than a spatially-structured crypt population. The grid state over time is displayed in colors, where crypts are colored by the genetic similarity based on neutral mutation patterns (principal components analysis reducing

100 neutral loci to three red, green, and blue dimensions, first row) and by the number of selective mutations they have acquired (magenta to red, second row). Clones expand locally driven by acquisition of fitness-increasing SGAs and neutral SGA hitchhike on the clonal expansions forming complex patterns according to crypts' phylogenetic histories.

We found that genetic diversity increased monotonically in 339 out of 360 simulation runs and decreased by ≥ 1% at any time during the remaining 21 simulation runs. We estimated "biopsy-based" genetic diversity over time by sampling 12 biopsies at random locations without replacement and successfully detecting mutations at neutral loci that are at ≥ 0% or ≥ 5% frequency in the population of 100 crypts comprising a biopsy in Figure 3.4 and Figure 3.5, respectively. We estimated "individual-crypt-based" genetic diversity over time by sampling 100 individual crypts without replacement at random locations within the grid in simulations with spatially-structured crypt population (Figure 3.2) and in simulations with fully-mixed crypt population (Figure 3.3).

3.5. Discussion

We assumed that neoplastic cell population in Barrett's esophagus is organized into a hexagonal 2D lattice and clones bearing acquired somatic genomic abnormalities (SGA) expand by crypt fission. The crypt cycle model was originally proposed by Totafurno et al. [94] and expansion of mutations by crypt fission was experimentally confirmed by Greaves et al. [95,96]. We hypothesized that BE crypts are not in a steady state, but are continuously cycling, given the observation of branching crypts in histopathology slides of biopsies sampled throughout segments [97]. In addition, we hypothesized that SGAs that occur in the self-renewing stem cell population at the bottom of crypts can expand spatially through successive cycles of crypt replication. This hypothesis is supported by evidence that the same mitochondrial DNA point mutations were found present in both arms of colon crypts undergoing fission, as well as in patches of neighboring crypts, thus establishing crypt fission as the mechanism for the spread of mutations in crypt-structured human colonic epithelium [95]. We assumed a population of 90,000 crypts in a typical Barrett's segment, which is not dramatically different from a mean 72,480 crypts per segment estimated from 13 individuals in Barrett's (Chapter 4, Supplementary table S1). Empirical evidence suggests that over the duration of endoscopic surveillance the Barrett's segment length remains constant over time (see Figure 1.2 in Chapter 1), therefore we held constant the width and height of the Barrett's segment during a simulation run.

In BE, typically a maximum of twelve biopsies are collected per endoscopic procedure, or time point during follow-up visit. We found that the underlying genetic diversity, which we estimated by comparing 100 individual crypts from random locations, was underestimated by this standard biopsy sampling protocol. This could be overcome if individual crypts are sampled from available biopsies.

Neoplastic cell populations can acquire SGAs, which can be produced by single catastrophic events, chromothripsis [98], or genomic firestorms [99], as well as a variety of other mechanisms including repair of double strand breaks by homologous recombination [100]. Some SGAs may increase neoplastic cell fitness resulting in clonal expansion that changes the frequencies of neutral and selective SGAs within and outside of the expanding clone (Figure 3.6). These changes in frequency modulate the level of genetic diversity and periodic measurements of genetic diversity provide a snapshot of evolutionary dynamics in the neoplasm.

The agent-based model we developed can be used to simulate a vast number of evolutionary scenarios and to calculate a vast number of summary statistics, such as number and sizes of clonal expansions and genetic diversity dynamics. The model allows the input of a user-defined fitness landscape by specifying the exact number of fitness-modulating loci and their associated fitness-modulating values. The model has a quick run-time and records a phylogenetic tree of the evolution of clonal crypt lineages and outputs complete information of the mutational state of the grid with associated statistics of clone sizes, parent relationships among clones, and fitnesses (replication rate and death rate) of clones. All of those capabilities allow the comparison of summary statistics of simulated data to experimentally-obtained data from individual Barrett's segments in future studies. Approximate Bayesian computation methods can assess distance between summary statistics of simulated and real data to estimate the underlying set of evolutionary dynamics parameter values, such as mutation rates, crypt expansion parameters, and number of fitness-modulating loci and their associated fitness modulating effects. Our model is an advancement in modeling evolutionary dynamics in its increased complexity by parameterizing genomic states of driver and passenger loci, fitness landscapes of driver loci, and by recording individual run phylogenies.

3.6. Conclusion

In summary, our findings suggest that genetic diversity most likely increases monotonically over time despite clonal expansions; and we predict that monotonic increases in genetic diversity over time would be observed *in vivo* in most individuals with BE. In such monotonic increase scenarios, measurements of genetic diversity at regular intervals can be used to estimate the elapsed time since initiation of Barrett's esophagus and to estimate the waiting time to progression to esophageal adenocarcinoma. The underlying genetic diversity is likely to be better estimated when genotyping individual crypts from biopsies as opposed to evaluating a mixture of crypts under most population parameter settings.

Chapter 4. NSAIDs modulate clonal evolution in Barrett's Esophagus

Authors: Rumen L. Kostadinov[1], Mary K. Kuhner[7], Carissa A. Sanchez[2,3], Patricia C. Galipeau[2,3], Thomas G. Paulson[2,3], Xiaohong Li[2,3], Cassandra L. Sather[4], Amitabh Srivastava[5], Robert D. Odze[5], Patricia L. Blount[2,3], Thomas L. Vaughan[3], Brian J. Reid[2,3,6,7], Carlo C. Maley[8]

Author affiliations: [1]Genomics and Computational Biology Graduate Program, University of Pennsylvania; Divisions of [2]Human Biology and [3]Public Health Sciences, Fred Hutchinson Cancer Research Center; [4]Genomics Resource, DNA Array Laboratory, Fred Hutchinson Cancer Research Center; [5]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Departments of [6]Medicine and [7]Genome Sciences, University of Washington; [8]Center for Evolution and Cancer, Helen Diller Family Comprehensive Cancer Center, Department of Surgery, University of California San Francisco;

Acknowledgements: We thank foremost the study participants who have made this study possible. We thank Dr. Junhyong Kim, and Dr. Paul Sniegowski for their helpful comments on the manuscript; Dr. Joseph Felsenstein for help with the mathematics for a binary evolutionary model; Jessica Arnaudo for contributions to analysis of segmentation results; David Cowan and Terri Watson for database support; Tricia Christopherson for NSAID interview management; Christine Karlsen and Valerie Cerera for patient care and research biopsy coordination.

Author contributions: RLK, MKK, CAS, PCG, TGP, XL, PLB, TLV, BJR, and CCM contributed to conception and design of the study. CAS and TGP isolated DNA from biopsies. CLS processed DNA with Illumina arrays. RDO and AS provided pathology diagnoses and analysis of crypts from histopathology slides. RLK, MKK, and CCM contributed to design of all computational methods. RLK implemented all computational methods and generated results and plots of the data. RLK, MKK, CAS, PCG, TGP, XL, PLB, TLV, BJR, and CCM contributed to analysis and interpretation of the data. PLB and BJR enrolled patients, and PLB collected the tissue samples. TLV collected and managed NSAID data. RLK drafted the manuscript. RLK, MKK, CAS, PCG, TGP, XL, CLS, AS, RDO, PLB, TLV, BJR, and CCM critically revised the manuscript.

Financial disclosure: Supported by NIH P01 CA91955, RC1 CA146973, R01 CA140657, R03 CA137811 and Research Scholar Grant #117209-RSG-09-163-01-CNE from the American Cancer Society

4.1. Abstract

Background

Cancer is considered an outcome of decades-long clonal evolution fueled by acquisition of somatic genomic abnormalities (SGAs). Non-steroidal anti-inflammatory drugs (NSAIDs) have been shown to reduce cancer risk, including risk of progression from Barrett's esophagus (BE) to esophageal adenocarcinoma (EA). However, the cancer chemopreventive mechanisms of NSAIDs are not fully understood. We hypothesized that NSAIDs modulate clonal evolution by reducing SGA acquisition rate.

Methods and Findings

We evaluated 13 individuals with BE. Eleven had not used NSAIDs for 6.2±3.5 (mean±standard deviation) years and then began using NSAIDs for 5.6±2.7 years, whereas two had used NSAIDs for 3.3±1.4 years and then discontinued use for 7.9±0.7 years. 161 BE biopsies , collected at 5-8 time points over 6.4-19 years, were analyzed using 1Million-SNP arrays to detect SGAs.

We observed a significant increase in the acquisition of new SGAs off-NSAIDs compared to on-NSAIDs (60±166 vs. 28±48 per biopsy, $p=0.013$) and in pre-existing SGAs that dropped out of detection on-NSAIDs as compared to off-NSAIDs (92±143 vs. 19±21 per biopsy, $p<0.01$). The estimated SGA rate was 7.8 per genome per year (95% support interval [SI], 7.1–8.6) off-NSAIDs and 0.6 (95% SI 0.3–1.5) on-NSAIDs. Twelve individuals did not progress to EA. In 10 we detected 279±86 SGAs affecting 53±30 Mb of the genome per biopsy per time point and in two we detected 1,463±375 SGAs affecting 180±100 Mb. In one individual who progressed to EA we detected a clone having 2,291±78 SGAs affecting 588±18 Mb of the genome at three time points in the last 3 of 11.4 years of follow-up.

Conclusions

NSAIDs were associated with reduced rate of acquisition of SGAs. The BE cell population maintained relative evolutionary stasis over prolonged periods in most individuals but occasionally stasis was punctuated by expansion of clones having massive amount of SGAs.

4.2. Introduction

Clonal evolution is a theory that explains the phenomenon of the progressive morphological and genetic change of somatic cell populations from normal homeostatic cell division and death within tissues to abnormal neoplastic growth and cancerous spatial expansion within and across tissues [2,27,101]. Clonal evolution is the Darwinian evolution by natural selection of asexually (mitotically) dividing somatic cells. Somatic genomic abnormalities (SGA), such as copy number alterations and loss of heterozygosity (LOH), can be used as polymorphic DNA markers for identifying evolving clones. Strictly defined, a clone is a genetically identical subpopulation of cells within the cell population of a tissue, that descends from a most recent common ancestor (MRCA) cell and therefore all of the clone's cells inherit the SGAs that were originally present in the MRCA cell. However, a commonly used, relaxed definition of a clone is descent with modification from a MRCA cell, which allows for accumulation of additional SGA heterogeneity among the cells of the clone. In other words, a clone ideally represents the shared cell lineage history of a subpopulation of cells. The acquisition of SGA variability (SGA polymorphism) over the course of cell division allows for classification of cell subpopulations into clones. In the remainder of this study, we use clone in its relaxed definition and we estimate phylogenetic trees from acquired SGA variability to qualitatively describe relatedness among evolving clones. The generation of new clones is stochastic and the change in clones' frequencies in the population is determined by clones' relative fitnesses. New adaptive and new neutral clones can arise stochastically over time [102] with every newly acquired SGA that does or does not affect fitness, respectively (Figure 1A,B). In order to prevent progression to cancer, mechanisms that modulate clonal evolution by either preventing SGA acquisition or preventing the spread of SGA-containing clones need to be elucidated.

Barrett's esophagus (BE) is a condition of the distal esophagus in which the normal stratified squamous epithelium is replaced by columnar epithelium with intestinal metaplasia [5]. BE is thought to develop as a complication of chronic gastroesophageal reflux disease (GERD), and individuals with BE are at increased risk of progression to esophageal adenocarcinoma (EA): 1-7 persons with BE progress to EA per 1000 person-years [6,7]. Strategies for early detection and prevention of esophageal adenocarcinoma have focused on all aspects of the GERD-BE-EA sequence: acid suppression medications, anti-reflux surgery, esophagectomy, ablation of premalignant Barrett's esophagus, endoscopic biopsy surveillance of Barrett's esophagus, and chemoprevention using aspirin or other non-steroidal anti-inflammatory drugs (NSAIDs) [5,103]. Barrett's esophagus is a pre-malignant condition in which clonal evolution can be studied *in vivo,* since a standard of care is periodic endoscopic surveillance with concomitant biopsy, providing a tissue bank that facilitates studies of clonal evolutionary dynamics over time.

Genomic instability is a common feature of solid cancers [2,29,36,38,104]. In a recent study, Beroukhim et al. evaluated 3131 cancer specimens from 26 histologic types and 1480 normal tissue specimens and found that copy number gains and losses affected 17% and 16% of the genome in a typical cancer specimen and only 0.35% and 0.1% of the genome in a typical normal tissue specimen [35]. Despite the recent massive accumulation of data on genomic alterations in cancers from the Cancer Genome Atlas and the International Cancer Genome Consortium initiatives, theoretical modeling of the generative process (clonal evolution) producing the observed SGA patterns and underlying neoplastic progression has remained limited [27,101]. BE is associated with genomic instability and acquired SGA [76,78,79,105] allowing analysis of the acquisition of SGA over time. This provides data for estimating SGA acquisition rate that is a key parameter of clonal evolution.

NSAID use significantly reduces the incidence and mortality rates of many types of cancer, including esophageal adenocarcinoma [64–68,106]. Rothwell et al. showed that the hazard ratio for cancer incidence of NSAID users vs. NSAID non-users was 0.66 (95% CI 0.50-0.87); however a robust NSAID cancer preventive effect manifests significantly only after ≥5 years of regular use [64]. The majority of epidemiological studies in BE suggest that NSAID use in individuals with BE reduces risk of developing EA [65–67,106]. Specifically, Vaughan et al. evaluated 350 individuals followed up for a median of 5.4 years (range 0.2-8.9) and showed that the 5-year cumulative incidence of EA was 14.3% (95% CI 9.3-21.6) for NSAID never users compared to 6.6% (3.1-13.6) for current NSAID users and that the hazard ratio for EA incidence of NSAID users vs. NSAID non-users was 0.20 (95% CI 0.10-0.41) [67]. Galipeau et al. showed that NSAID use reduced the 10-year cumulative incidence of esophageal adenocarcinoma from 79% to 30% in individuals with Barrett's esophagus who had one or more somatic genomic abnormalities detected at baseline endoscopy, which included DNA content tetraploidy and/or aneuploidy, assayed by DNA content flow cytometry, or genetic abnormalities, such as loss of heterozygosity (LOH) on chromosomes 9p and 17p, assayed by PCR of small tandem repeat (STR) loci [68]. NSAID use for chemoprevention is attractive due to the widespread use and low toxicity and side effects of that class of drugs; however the molecular mechanisms underlying the NSAID cancer preventive effect are not fully understood. In this study, our aim was to evaluate the effect of NSAIDs on the accumulation of somatic genomic abnormalities by evaluating the entire genome (1 Million SNP loci) for SGA. We hypothesized that NSAID use modulates clonal evolution by reducing the prevalence of SGA by either reducing the incidence of SGA over time (SGA rate: number of SGAs acquired per genome per year) or interfering with the expansion of clones bearing newly acquired SGAs over time (Figure 1A,B).

To test this hypothesis, we used a prospective observational crossover study design: a longitudinal study in which the sequence of NSAID use was recorded for each individual during the follow-up period. We selected thirteen individuals with BE from our cohort, who had endoscopic follow-up of mean 11.8 ± 3 years (range: 6.4–19) and who began or discontinued NSAID use exactly once during follow-up. All thirteen individuals had to have at least two consecutive time points (≥6 biopsies) off NSAIDs and at least two consecutive time points (an additional ≥6 biopsies) on NSAIDs. To estimate SGA prevalence in biopsies on and off NSAIDs we used summary statistics of observed patterns of SGA; to estimate SGA rates on and off NSAIDs we used an evolutionary analysis of observed SGA patterns to take into account SGA phylogenetic identity by descent. Drummond et al. showed that mutation rates can be estimated from longitudinal samples in virus populations using coalescent and phylogenetic methods within a Bayesian Markov Chain Monte Carlo framework for sampling model parameter space (BEAST package, Bayesian Evolutionary Analysis Sampling Trees) [62,107]. We adapted BEAST to estimate SGA acquisition rates on and off NSAIDs. Thus, the crossover study design provided 13 independent tests of the hypothesis of NSAID-associated reduction in SGA acquisition rate since every individual had both on and off NSAID periods and SGA acquisition during those periods.

4.3. Results

We evaluated the dynamics of detected SGAs over time. The mean number of SGAs and the proportion of the genome they affected remained approximately constant over time, for as many as 19 years (e.g., Figure 4.2, individual a). Individuals b, f, and j, shown in red in Figure 4.2A and Figure 4.2B, showed much greater variation in detected SGA over time, compared to the rest of the individuals, shown in black. Progression to EA was not part of our study inclusion criteria, and individual j was the only individual who progressed. Individual f did not progress to EA, but rather opted for esophagectomy for high-grade dysplasia after 6.4 years of follow-up and subsequently died of a different cancer 11.9 years later. In individuals b,f, and j, the mean (± standard deviation) number of SGAs per genome per time point was 1,082 ± 177 , 1,844 ± 573, and 1,154 ± 746 respectively, and the amount of genome affected by SGAs was 119 ± 79 Mb, 242 ± 121 Mb, and 227 ± 222 Mb, respectively. In the rest of the individuals, the mean number of SGAs per genome per time point was 279 ± 86 and the amount of genome affected was 53 ± 30 Mb. Assuming a human genome length of 3,164 Mb (Human genome GCRh37.p5 assembly), individuals b, f, and j had 3.8 ± 2.5%, 7.6 ± 3.8%, and 7.2 ± 7% altered somatic genome per time point, compared to 1.7 ± 0.9% altered somatic genome in the rest of the individuals. Figure 4.2 suggests an overall evolutionary stasis in this sample of persons with BE, i.e., in most individuals, contrary to expectations, we did not observe a gradual increase of SGA over time.

The unexpected result of long-term evolutionary stasis suggested that the effect of NSAIDs on reducing SGA rate would be challenging to detect. We evaluated newly-appearing SGAs during periods of NSAID use and NSAID non-use by excluding all SGAs that were detected at baseline. Baseline SGAs were excluded since they had occurred and increased in frequency for an unknown amount of time prior to detection at baseline and since we have self-reported NSAID use information reaching back only 6 months prior to baseline. Across all individuals, we detected 28 ± 48 newly appearing SGAs per biopsy during on-NSAID periods (n=73 biopsies) compared to 60 ± 166 newly appearing SGAs per biopsy during off-NSAID periods (n=57 biopsies), which was a significant difference (Wilcoxon test, p=0.013). We also evaluated whether NSAID use is associated with regression of pre-existing SGAs (dropping out of detection), by evaluating only SGAs that are detected at baseline, or during NSAID use or non-use periods, but not detected in the last endoscopy. A significantly greater number of existing SGA events dropped out of detection on-NSAIDs, as compared to off-NSAIDs (92 ± 143 [n=55] vs. 19 ± 21 [n=79], Wilcoxon test, p<0.01). In summary, NSAID use was associated with a reduced number of newly appearing SGAs of any size (Figure 4.3A) and with higher number of pre-existing SGAs of any size dropping out of detection (Figure 4.3B).

While the natural history of clonal evolution was different in each individual, some common patterns can be discerned. The majority of SGAs are present in the first time point, with little accumulation of SGAs afterwards (Figure 4.6). This can also be seen in the radial spokes apparent in the Circos plots that summarize SGAs across the whole genome at each time point (Figure 4.4B, Figure 4.4C, Figure 4.5B, Figure 4.5C, and Supplementary Figure 4.7), that represent the most common lesions in BE [76,78], on chromosomes 9p (*CDKN2A*), 3p (*FHIT*) and 16p (*WWOX*). We found no evidence of selective sweeps of clones throughout the Barrett's segment in any of our 13 patients over at total of 153 patient-years. This can be seen in the consensus trees generated by BEAST (Figures Figure 4.4D, Figure 4.4G, Figure 4.5D, Figure 4.5G, and Supplementary Figure 4.8). These genealogies show that multiple clones appear to co-exist over the entire period of follow-up. Even in the one patient who progressed to EA, individual j, the clone with massive SGAs (2,291 ± 78 SGAs affecting 588 ± 18 Mb or 19% of the genome in biopsies 8,10,11, and 13) remained localized (Figure 4.4). Interestingly, a precursor of that clone had been detected 9 years prior to its emergence (biopsy 2 in Figure 4.4G,H,I). We show clonal evolution in individuals b, j, and f in higher detail in figures 4 and 5 since these individuals had a higher than average number of SGA events and amount of genome altered (Figure 4.2). We show clonal evolution in individual l (Figure 4.5) in higher detail to show clonal evolution during an on-off NSAID use pattern. In addition, the SGA amount in individual l is close to the mean SGA amount in all individuals, except b,f, and j, while SGA amount in individual f is higher than the mean (Figure 4.2) and using Circos plots side-by-side contrasts qualitatively SGA amount and SGA chromosomal location between individuals l and f (Figure 4.5 B,C). In summary, the majority of patients showed little gradual accumulation of new SGAs consistent with long-term evolutionary stasis during follow-up (Circos plots in Figure 4.4, Figure 4.5 and Supplementary Figure 4.7), and the one progressor to EA, individual j, showed that evolutionary stasis can be punctuated by the expansion of a clone with massive amount of SGAs (Figure 4.4 C,G,H,I).

The maximum-parsimony phylogenetic analysis revealed the shared common ancestry of biopsies within an individual based on SGA homology. Inferred PAUP* phylogenetic trees, which had branches scaled by the estimated number of shared SGA events in Figure 4.4E,H, Figure 4.5E,H, and Supplementary Figure 4.9, showed significantly imbalanced tree shapes (Supplementary Table 4.3) for all individuals, except individual f and j. When we rescaled the branch lengths of the same phylogenetic trees by the amount of genome affected in Figure 4.4F,I, Figure 4.5F,I, and Supplementary Figure 4.9, the trees showed that within an individual the majority of biopsies cluster together and only few biopsies or lineages shoot out of the majority cluster, which is indicative of SGA bursts.

We tested our hypothesis that NSAID use reduces SGA acquisition rate in Barrett's esophagus by estimating SGA acquisition rate during off-NSAID and on-NSAID periods using a custom modified version of BEAST. We added a new evolutionary model of SGA into BEAST in order to estimate SGA rate using Bayesian MCMC sampling (see Methods and Equations S3,4). We excluded SGAs detected in the first time point and only measured SGAs that were detected during follow-up, in order to reduce the influence of clonal evolution that occurred prior to surveillance. For the two individuals who were already on NSAIDs when we started surveying them, and later went off NSAIDs, individual l showed a lower SGA rate on NSAIDs than off NSAIDs, but the 95% support intervals for the two rates overlap (Figure 4.6). In contrast, individual m showed a higher SGA rate on NSAIDs than off NSAIDs. In individuals a-k, the SGA rate on NSAIDs was approximately an order of magnitude lower than the SGA rate off NSAIDs, with non-overlapping 95% support intervals (Figure 4.6), which is consistent with the hypothesis that NSAID use reduces SGA acquisition rate (on average 7.8 SGAs per genome per year off-NSAID vs. on average 0.6 SGAs per genome per year on-NSAID in individuals a-k).

4.4. Discussion

Somatic genomic abnormalities inevitably occur in a population of asexually (mitotically) dividing somatic cells, and if such SGA affect cell fitness, the cell population will evolve by natural selection and may evolve neoplastic and cancerous overgrowth. It is virtually unknown how clonal evolution unfolds in a human neoplasm over time since long-term observation is not feasible in the majority of benign and malignant conditions. However, in this longitudinal study of the premalignant condition Barrett's esophagus, we were able to evaluate the evolutionary dynamics of SGA in neoplasms of 13 individuals over more than a decade of follow-up, in 5-8 time points, and estimate the effect of NSAID use on modulating SGA dynamics.

We observed long-term evolutionary stasis, detecting an approximately unchanging mean number of SGAs across multiple biopsies and across multiple time points over an average of 11.6 years of follow-up (Figure 4.2). The estimated SGA rate on-NSAID and off-NSAID was 0.6 and 7.8 SGAs per genome per year, respectively (Figure 4.6). Few newly-appearing SGAs were observed during follow-up, either off-NSAID or on-NSAID periods (Figure 4.3A, Supplementary Figure 4.12). The maximum-parsimony phylogenetic analyses show tree topologies consistent with gradual accumulation of SGA events on the leading-edge of a clone, possibly during initiation that results in an imbalanced tree shape (Figure 4.4E, H, Figure 4.5E, H, and Supplementary Figure 4.9 and Supplementary Table 4.3). However, when inferred SGA events were converted to amount of genome they affected, the branch lengths (x-axis) of trees show little gradual accumulation of SGA amount (Figure 4.4F,I, Figure 4.5F,I, and Supplementary Figure 4.10) consistent with relative evolutionary stasis. These two results suggest that the accumulation of SGA events only affected a tiny fraction of the genome. The observation of long-term evolutionary stasis is consistent with a hypothesis that BE can function as a benign and perhaps protective evolutionary adaptation of epithelial tissue to duodenal gastroesophageal reflux [8]. According to that hypothesis, the multilayer squamous population of cells at the distal end of the esophagus encounters a new, acidic microenvironment when the development of a hiatal hernia, often associated with BE, permits chronic exposure to reflux. Constitutive germline variants (evolution at the population level) can synergize with acquired somatic genomic abnormalities (evolution at the somatic tissue level) to modulate the propensity for developing BE that appears to be a successful adaptation of the tissue in response to exposure to reflux [8]. In addition, the BE epithelium may have a survival advantage over the native multilayer squamous because it is a columnar secretory epithelium that expresses mucosal defense functions, such as bicarbonate secretion [16], mucus secretion by goblet and other columnar cells [18], expression of claudin-18 tight junctions [23], and overexpression of genes involved in defense and repair [15,24,108,109]. The observation of evolutionary long-term stasis at the genome level is also consistent with the hypothesis that the BE is a benign condition that rarely progresses to EA, corroborated by epidemiological evidence that only 1-7 persons with BE progress to EA per 1000 person-years [6,7].

However, apparent evolutionary stasis at the level of analyses of biopsies may miss ongoing accumulation of SGAs within single crypts. If those clones never grow larger than a few crypts, they would not be detected by our assays. Further work will be necessary to determine if the stasis seen at the biopsy level is a result of the lack of accumulation of SGAs in crypts or the lack of clonal expansions of those SGAs to detectable sizes. Our selection criteria, both for patients that have used NSAIDs, and for at least 5 time points over at least 6.4 years of follow-up may have led to selection bias for individuals with relative evolutionary stasis in their Barrett's epithelium.

While evolutionary stasis dominated the dynamics of SGA over time, NSAID use was associated with detectable reduction in the rate of acquisition of SGA. NSAID use reduced SGA rate from an average of 7.8 SGAs per genome per year off-NSAID to 0.6 SGAs per genome per year on-NSAID (Figure 4.6). NSAID use was also associated with reduction in the number of newly appearing SGAs and increase in the number of pre-existing SGAs dropping out of detection (Figure 4.3, Supplementary Figure 4.12). These results suggest that a possible underlying mechanism of the NSAID effect on reducing cancer risk may be modulating SGA evolution by reducing the acquisition of new SGA, inhibition of spread of SGA-containing clones, preventing massive bursts of SGA, and possibly induction of apoptosis in clones having high numbers of pre-existing SGAs.

Evolutionary stasis can be punctuated by a massive catastrophic burst of SGA. Individual j was the only individual who progressed to EA and the only one that showed the sudden appearance of a clone with massive SGA affecting 19% of the genome, yet that clone remained stable for ~3 years prior to EA occurrence (Figure 4.4 A,C,G,H,I). Interestingly, biopsy 2, taken at the baseline endoscopy 8.5 years prior to detecting the clone with massively altered genome, shared a subset of the SGAs with that clone (chromosomes 10, 12, 17 and 18), and thus biopsy 2 is likely to be an early, ancestral progenitor of the massively altered clone. This one observation illustrates the need for larger studies to identify altered states of the genome, early in progression, that may prime the genome for future bursts of massive alterations. We hypothesize four possible mechanisms for a long-term dynamic of stasis interrupted by a massive burst and followed again by stasis: 1) small SGAs or point mutations that we miss by our assay may accumulate and hit key genes involved in DNA maintenance and repair, leading to a massive catastrophic event that disrupts the genome, 2) massive bursts of SGA occur often, but in most cases bursts have deleterious consequences on cell fitness and those clones do not survive, 3) epistatic interactions among multiple loci in the genome exist in such configuration that makes acquisition of SGAs at any one of the loci deleterious, but simultaneous acquisition of SGAs at multiple loci beneficial and 4) SGAs gradually accumulate within a small region of the Barrett's segment, potentially as small as a single crypt, that we miss in biopsy sampling until they eventually expand. The first hypothesis might be supported or ruled out with more detailed data using next-generation sequencing. However, it may not be likely since we did not detect apparent continued accumulation of SGAs over time in individual j; if DNA maintenance or repair mechanisms were inactivated we would expect a linear or exponential increase in the amount of SGA after the burst. It is also possible that SGA continued to accumulate in cells descended from the original burst, but these subclones did not expand to detectability, so these mutations were not visible in our assay. The second hypothesis may be likely given the rare observation of biopsies having a burst in SGA, for example biopsy 5 in individual b (Figure 4.4 A,B,D,F). The third hypothesis can only be supported or ruled out by a cross-sectional study with hundreds of between- and within- individual samples that might reveal a signature of co-occurring SGAs. The fourth hypothesis can be evaluated by analyses of the SGA heterogeneity among individual crypts within a biopsy and analyses of the relationship between spatial distance and genetic distance among biopsies and crypts within the Barrett's segment. These four hypotheses are not mutually exclusive and may instead synergize to generate the observed stasis-punctuation-stasis evolutionary dynamic. Interestingly, the relative stasis observed after the massive SGA punctuation event is consistent with the observation of a relatively stable aneuploid population of cells detected in a primary breast tumor and its associated liver metastasis [110]. Moreover, the pattern of massive SGA (Figure 4.4C) may have been generated by a multipolar mitosis [29] or a sequence of bridge-fusion-breakage cycles and could be classified as chromosomal instability, "chromothripsis", "complex genomic firestorm", or genetic catastrophe types of SGA patterns that have been observed in other solid tumors [35,36,38,98,99,110].

4.5. Methods

Human Subjects

Participants were selected from the Seattle Barrett's Esophagus Study, a research cohort founded in 1983. Surveillance endoscopies were performed and biopsies were taken using a standardized four quadrant sampling protocol [111]. At endoscopy, anatomical landmarks including the gastroesophageal junction (GEJ) and *ora serrata* (OS) were noted, which define the lower (distal) and upper (proximal) boundaries, respectively, of the Barrett's segment. During an endoscopy, biopsies were taken every one or two cm along the length of the Barrett's segment. At each level, four biopsies were taken approximately at 0°, 90°, 180°, and 270° around the circumference of the esophagus for histologic evaluation. Endoscopic biopsies for molecular studies were collected in Minimal Essential Media (MEM) with 10% DMSO (Sigma #D-5879), 5% heat inactivated fetal calf serum, 5mM Hepes buffer on ice and frozen at -70°C. In 1995 the research protocol added an epidemiologic interview in which individuals were questioned about NSAID use, as previously described [67]. In addition, the protocol added blood collection at the time of endoscopy for use as a control, since blood DNA represents putatively unaltered germline genotype.

Study design

Individuals were selected in the cohort who had at least a 3cm-long BE segment at baseline. Individuals were further selected based on NSAID use status changing exactly once during prospective follow-up and based on having at least two endoscopic procedures while using NSAIDs and at least two while not using NSAIDs. At least five years of follow-up was also required in order to observe evolution over time. Thirteen individuals met these inclusion criteria (Figure 4.1C). The history of NSAID use at each endoscopy was evaluated with a questionnaire that was also used in a US collaborative case-control study of esophageal adenocarcinoma [112]. As part of the questionnaire, individuals are shown cards (i.e., typed lists of drugs with trade names and generic names) to facilitate recall. Individuals were also asked about indications for taking NSAIDs, and reasons for stopping in those who were no longer regular users. The criterion for regular NSAID use at an endoscopy was taking an NSAID at least once per week for the last 6 months. Regular NSAID use over multiple endoscopies defines a time interval on-NSAIDs and absence of NSAID use over multiple endoscopies defines a time interval off-NSAIDs. We approximated the transition point between NSAID use and non-use by taking the middle time point equidistant between the two endoscopies when the NSAID use changed (Figure 4.1C, white-gray boundary). Eleven individuals (a-k) were not on NSAIDs at the start of surveillance and then went on NSAIDs (had an "off–on NSAIDs" pattern during surveillance), and two individuals (l,m) had the opposite, starting surveillance on NSAIDs and then stopping their use (an "on–off NSAIDs" pattern). The median follow-up surveillance time per individual was 11.6 years (range 6.3-19). A total of 74 endoscopies and 161 biopsies were selected as well as one blood sample for each of the thirteen individuals to serve as normal constitutive genotype control.

Sample preparation

The 161 frozen biopsies were thawed and rinsed in Hanks buffered salt solution without divalent cations (HBSS, Gibco/BRL). Biopsies were incubated 60 minutes at room temperature in 30mM EDTA in HBSS preheated to 37°C. Barrett's epithelium was isolated by gently peeling it away from the stroma with microdissection needles under a dissecting microscope [73]. The 13 frozen blood samples were processed the same way as the biopsies, except for the epithelial isolation step. DNA was extracted using Puregene DNA Isolation Kit as recommended by the manufacturer (Gentra Systems, Inc. Minneapolis, MN). Samples were quantitated using the Picogreen method (Quant-iT dsDNA Assay, Invitrogen). A total of 200ng of DNA at 50ng/ul concentration was analyzed using Illumina Omni-Quad 1M SNP arrays according to manufacturer's protocol. All samples were evaluated at the Fred Hutchinson Cancer Research Center Genomics Core Laboratory.

GenomeStudio processing

All raw intensity files were loaded in Illumina's GenomeStudio v3, normalized and clustered using the SNP manifest and cluster files for build37 of the human genome. In all our analyses we used the total signal intensity R for each SNP, which is the sum of the normalized X ("A" allele, Cy5 red) and Y("B" allele, Cy3 green) intensities. We also used the B allele frequency (BAF), which is a modified version of the allelic intensity ratio theta (θ = 2/p*arctan(Y/X)), to reduce SNP-to-SNP variation in theta using the canonical clusters.

GLAD segmentation

Each individual's BE DNA samples were paired to the individual's control sample (DNA from blood from the same individual), which always appeared normal, i.e. lacking any chromosomal alterations (none of the control samples had any split in BAF over the entire genome, data not shown). For each individual, we first excluded the 0.2% of SNPs with the lowest R values in the control sample, to remove SNP probes that either perform poorly or fall within germline copy number variant (CNV) regions. We corrected for dye bias (higher fluorescence of the B allele, Cy3 green) by re-centering the BAF of heterozygous and homozygous SNPs of all samples from observing that the median BAF of heterozygous

SNPs was ~0.53, instead of 0.5. Then, for each individual, we identified the set of heterozygous SNPs; i.e., SNPs having a BAF in control sample between 0.33 and 0.66. Finally, we separated the data into three signal profiles: $\log_2$ (R of BE sample / R of control sample) for heterozygous SNPs only, $\log_2$ (R of BE sample / R of reference) for homozygous SNPs only, and reflected and scaled BAF of BE sample, (mBAF= abs(BAF of BE sample – 0.5)*2) for heterozygous (informative for LOH) SNPs only. We performed separate wavelet-based segmentation on these three signal profiles using the HaarSeg algorithm [113] from the GLAD [80] package (using parameters haarStart=3, haarEnd=9, fdrQ=0.0001, onlySmoothing=T).

SGA detection

For each biopsy sample, we combined all break points of the segmented three signal profiles to create a new set of segments that is the union of the segments from the three signal profiles unique to each individual's biopsy. For every new segment, we used thresholds to call allelic imbalance based on the smoothed mBAF profile, and to call single or double copy gain or loss, based on the homozygous and heterozygous $\log_2$R profiles. Thus every new segment meeting the thresholds received one of eight molecular state calls: AB (normal), AA (copy neutral LOH), A (single copy loss), 0 (double copy loss), AAB (single copy gain), AAA (LOH plus subsequent single copy gain), AAAA (LOH plus subsequent double copy gain), AABB (double copy gain). In summary, Supplementary Table 4.2 shows all calling thresholds used and supplementary figures showing raw data segmentation and SGA calls for all 161 biopsies of individuals a-m.

The GLAD segmentation detects break points of SGA for each sample individually. For each individual, we combined all break points across the individual's samples and ran a segment merging procedure that merged two adjacent, neighboring segments if they had the same molecular state call across all samples of that participant. Thus, the number of segments per individual can vary. IMPUTE2 [114] and a reference dataset of 566 CEU haplotypes, part of the 1000 Genomes Project [115], was used to phase each individual's blood control sample. Having haplotype assignments for the A and B alleles of every SNP, we developed an algorithm to assign a haplotype state for every segment of allelic imbalance. This results in conversion of AA, A, AAB, AAA, AAAA calls to BB, B, BBA, BBB, BBBB calls for segments having lost or gained the opposite allele. For simplicity of all subsequent analyses, all segments having AB molecular states were assigned an "absence of SGA" call, and all segments having other molecular states were assigned a "presence of SGA" call. The final results are individual-specific phylogenetic matrices having samples as taxa, chromosomal segments as characters, and binary molecular states (SGA absence/presence, or 0/1) as character states.

Phylogenetic analyses

To measure mutation rate change associated with NSAID use, we used a two epoch model in BEAST [107], where the transition time between the first and second sampling periods is the time of change in NSAID use. We ran BEAST for 10 million Bayesian MCMC iterations that sample the space of genealogies and population parameters to obtain posterior distributions for model parameters that best fit the data. We used uniform prior distributions for SGA rate with lower and upper bounds of 10^{-5} and 10^{4} SGAs per biopsy genome per year, respectively, for the duration of any of the on-NSAID and off-NSAID periods and estimated SGA rate separately for the first and second sampling periods, where each SGA rate adheres to the molecular clock hypothesis (SGA occur at constant rate for all evolving lineages) for the period duration. We added a 0/1 mutation model in BEAST for the SGA absence/presence character states (see Equations S3-4) and this model assumed that SGAs do not revert to the normal type, i.e., 1→0 transition is impossible. We also modified BEAST's likelihood calculation algorithm to consider a last universal common ancestor (LUCA) that has an unaltered genomic state (zeros for all sites), and that connects to the most recent common ancestor (MRCA), at the root of the tree, creating an extra LUCA-MRCA branch. Thus, the final likelihood of the tree is the product of the likelihood of the tree at the root, calculated with Felsenstein's pruning algorithm [116], multiplied by the probability of the LUCA-MRCA branch length. Maximum parsimony trees were estimated using Wagner parsimony with delayed transformation (DELTRAN) on the individual-specific phylogenetic matrix with 0/1 SGA states using the PAUP* program [117]. For PAUP* analyses, we also used a character transition matrix that assumes infinite cost for 1→0 transitions, i.e. SGAs do not revert to normal type.

4.6. Figures

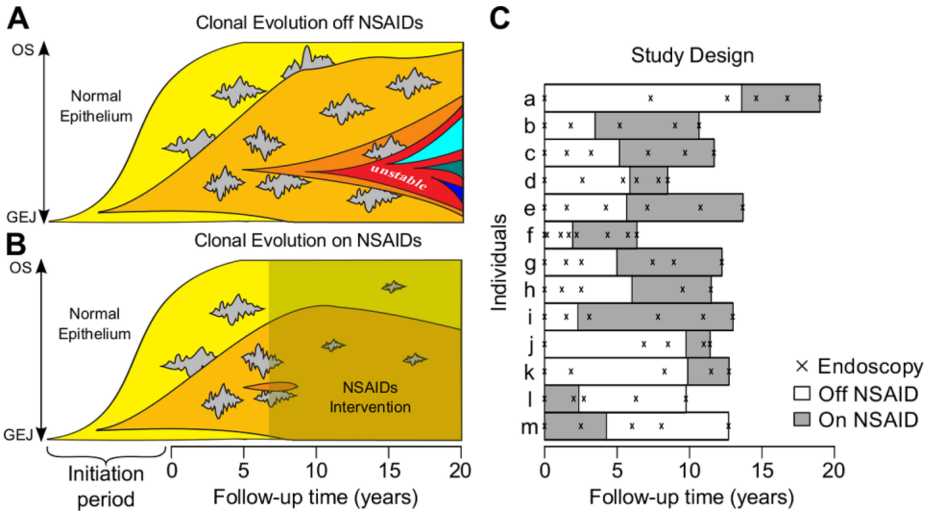


Figure 4.1. Hypothesis and study design for evaluating the NSAID effect on clonal evolution in Barrett's esophagus (BE). (Panels A, B) BE is a condition in which the normal stratified squamous epithelium (white) of the distal esophagus is replaced by specialized intestinal metaplasia (colors). During endoscopic surveillance (0-20 years, x-axis), the anatomical landmarks gastroesophageal junction (GEJ) and "ora serrata" (OS) define the lower (distal) and upper (proximal) boundaries of the Barrett's segment (y-axis), respectively. The mean distance between GEJ and OS is 5 cm in our cohort, and typically remains constant in size throughout 0-20 years of follow-up time. While the origin and initiation of BE is debated, we followed the model of Wang et al. where BE is thought to arise at GEJ from a residual embryonic population [118]. This initiation model is also consistent with observations of a columnar, secretory epithelium that forms superficial esophageal glands before being displaced by stratified squamous epithelium during embryonic development [119]. We estimated that the mean length of the initiation period in BE is 5.81 years by measuring crypt density and fraction of branching crypts and assuming a single progenitor crypt and logistic population growth of crypts by crypt fission (Supplementary Table 4.1). Somatic genomic abnormalities (SGA) that confer a selective advantage give rise to clones that increase in frequency in the neoplasm over time (adaptive SGAs, yellow to blue colors). SGA that are selectively neutral give rise to clones that fluctuate in frequency in the neoplasm over time by genetic drift (neutral SGAs, gray). (Panel A) In the absence of NSAID use, clonal evolution is fueled by acquisition of SGA. Chromosomal instability (red unstable clone) can lead to increased clonal genetic diversity and progression to cancer. (Panel B) We hypothesized that long-term NSAID use lowers the rate of SGA acquisition. (Panel C) To test this hypothesis, we evaluated 13 individuals with BE, eleven of whom were not using NSAIDs (off-NSAIDs) for 6.2 ± 3.5 years (mean ± standard deviation) and then began using NSAIDs for 5.6 ± 2.7 years, and two of whom were using NSAIDs for 3.3 ± 1.4 years and then discontinued use for 7.9 ± 0.7 years. Frozen biopsies were assayed from 5–8 endoscopies

from each individual, marked with x's. The DNA from 161 BE biopsies and 13 blood samples was analyzed using 1M SNP arrays to detect SGA.

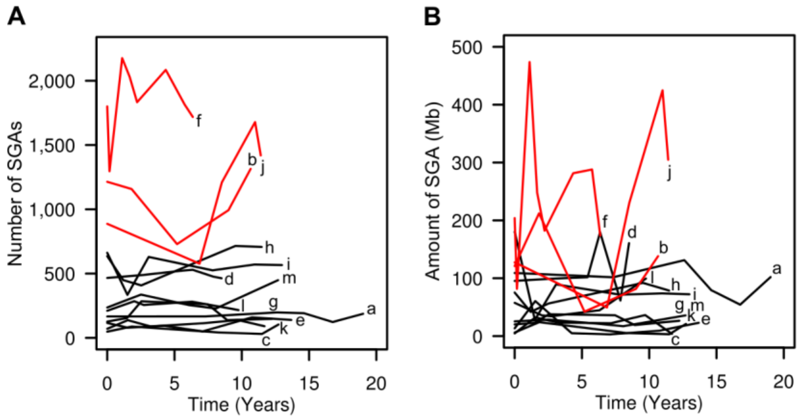


Figure 4.2. The mean number of detected somatic genomic abnormalities (SGA) and the amount of the genome they affect remain approximately constant over time in Barrett's esophagus. (Panel A) Solid lines connect the means at each time point for all individuals (a-m), where the symbols a-m are plotted at the end of the lines. The mean number of SGAs per biopsy-genome per time point was 1,082 ± 177, 1,844 ± 573, and 1,154 ± 746 (mean ± standard deviation) in individuals b, f, and j, respectively, compared to 279 ± 86 in the rest of the individuals. In this instance, number of SGAs is an individual-specific estimate of the total number of independently acquired SGA events and is computed by counting the number of abnormal genomic segments identified by the union set of all detected SGA break points across samples of a given individual (See *GLAD segmentation and SGA detection* in Methods). (Panel B) The mean amount of genome affected by SGA per time point was 119 ± 79 Mb, 242 ± 121 Mb, and 227 ± 222 Mb for individuals b, f, and j, and 53 ± 30 Mb for the rest of the individuals.

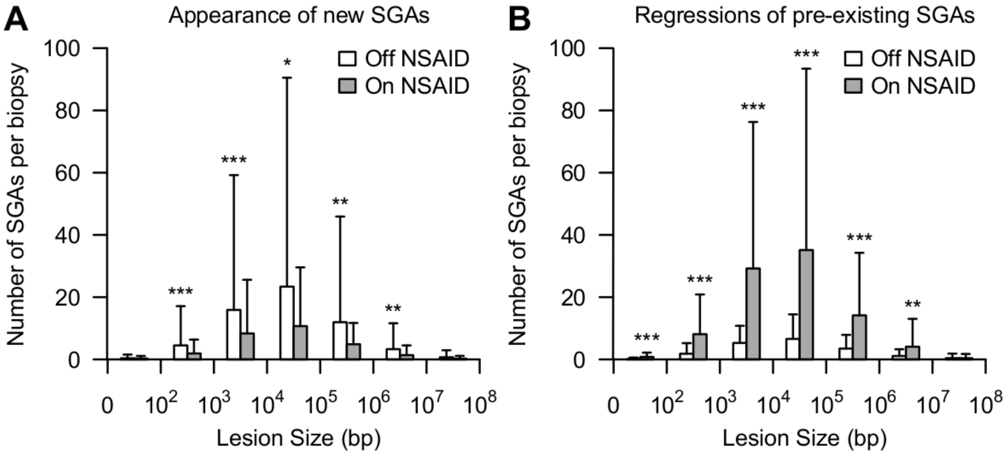


Figure 4.3. The effect of NSAIDs on appearance and regression of SGA events. (Panel A) NSAID use is associated with appearance of fewer new SGA events. For this analysis, we excluded all SGAs present at baseline because they had occurred and increased in frequency for an unknown amount of time prior to detection at baseline and since we have self-reported NSAID use information reaching back only 6 months prior to baseline. We counted only new SGAs that appeared within the off-NSAID or on-NSAID periods. (Panel B) NSAID use is associated with a decrease in detectable cell populations with pre-existing SGAs. For this analysis, we restricted the analysis to only the SGAs not detected in the final endoscopy in order to count their regression during either the on-NSAIDs or off-NSAIDs periods. (Panels A, B) We binned newly appearing or regressed pre-existing SGA according to lesion size (0 bp–100Mb, x-axis), but detected no apparent effect of NSAID use on selection for or against lesions of a specific size category; rather, NSAID use affected all size categories of SGAs equally. (Wilcoxon rank-sum test, 2-sided p-values, *** p<0.01, ** p<0.05, * p<0.1, solid bars and associated error bars represent mean and standard deviation of newly appearing and regressing SGAs per biopsy).

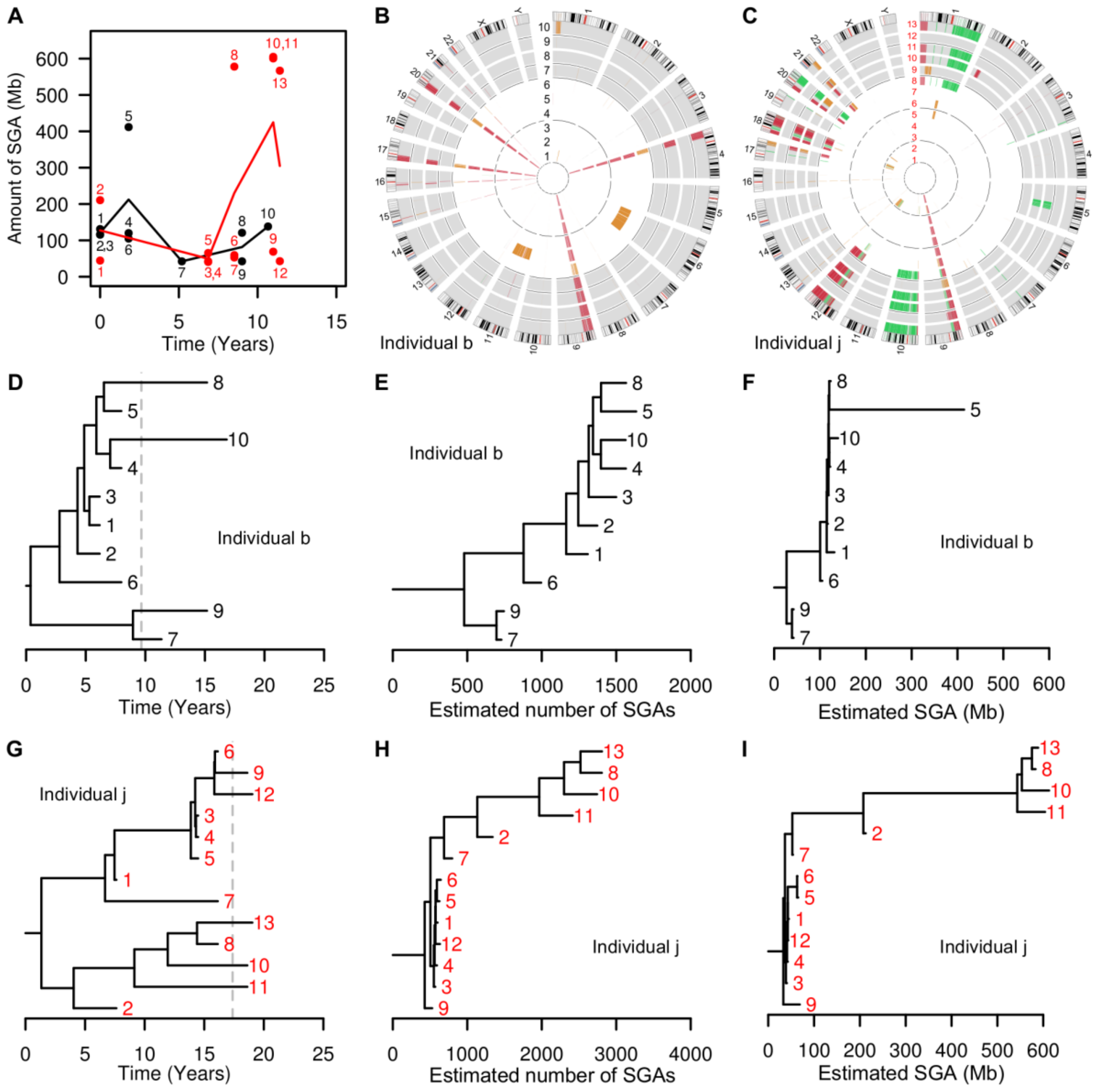


Figure 4.4. Clonal evolution in individuals b and j. (Panel A) Solid lines connect the mean amount of SGA detected across biopsies at each time point. Dots correspond to biopsies taken during follow-up (x-axis) that have total SGA detected by SNP arrays (y-axis). In individual b (black line), we observed evolutionary stasis, where the mean amount of SGA was 119 ± 79 Mb over more than a decade of follow-up. In individual j (red line), we observed evolutionary stasis up to year 7, which was disrupted by a massive burst of SGA detected in year 8.5. Three years after the appearance of this massively altered clone, individual j progressed to esophageal adenocarcinoma. Individual b started NSAIDs after year 5, while individual j started regular NSAIDs use only after year 10. (Panels B and C) Genome-wide view of SGA over time in individuals b and j. Each ring, labeled with a biopsy number, represents whole-genome SGA data from a different biopsy. Thin black line rings separate endoscopies (time points), white background shows time periods off-NSAIDs and

gray background shows time periods on-NSAIDs. Within the rings, black segments designate homozygous deletion, red single copy loss, orange copy-neutral LOH, and green shows copy gain. (Panel B) Circos plot of SGA in individual b. Note the appearance of "new" whole chromosome LOH at chromosome 6 and 11 in biopsy 5, taken during the off-NSAIDs period, and the detection of a clone lacking alterations on chromosomes 4, 12, 17 and 20, in biopsies 9 and 7, taken during the on-NSAIDs period. (Panel C) Circos plot of SGA in individual j. A massive burst of SGAs was detected first in biopsy 8, in year 8.5, before the individual began regular NSAID use. Biopsy 2 (second inner ring), taken at the baseline endoscopy 8.5 years prior to the burst, shared a subset of the SGAs seen in the massively altered clone (chromosomes 10, 12, 17 and 18), and thus is likely to be an early, ancestral progenitor of the massively altered clone. (Panels D and G) Consensus phylogenetic trees estimated by BEAST reveal long-term co-existence of clones. Branch lengths are scaled according to time, the tips of the phylogeny are biopsies aligned on the x-axis according to their sampling time, and all internal nodes are estimated coalescence times assuming a logistic population growth model (see Methods and Text S1). Dashed gray line represents the onset of NSAID use. In participant j, we detected 1,215 SGAs affecting 211 Mb of the genome in biopsy 2, the likely progenitor clone that presaged the appearance of 2,357 SGAs affecting 578 Mb of the genome in biopsy 8, 8.5 years later. In participant b, biopsies 7–10 have few SGAs and only a small amount of genome affected by SGA. (Panels E, F, H, I) Maximum parsimony trees estimated by PAUP reveal the ancestral relationships among biopsies based on shared SGA characters. Branch lengths are scaled according to estimated number of SGAs (Panels E, H) or the amount of genome affected by SGA (Panels F, I).

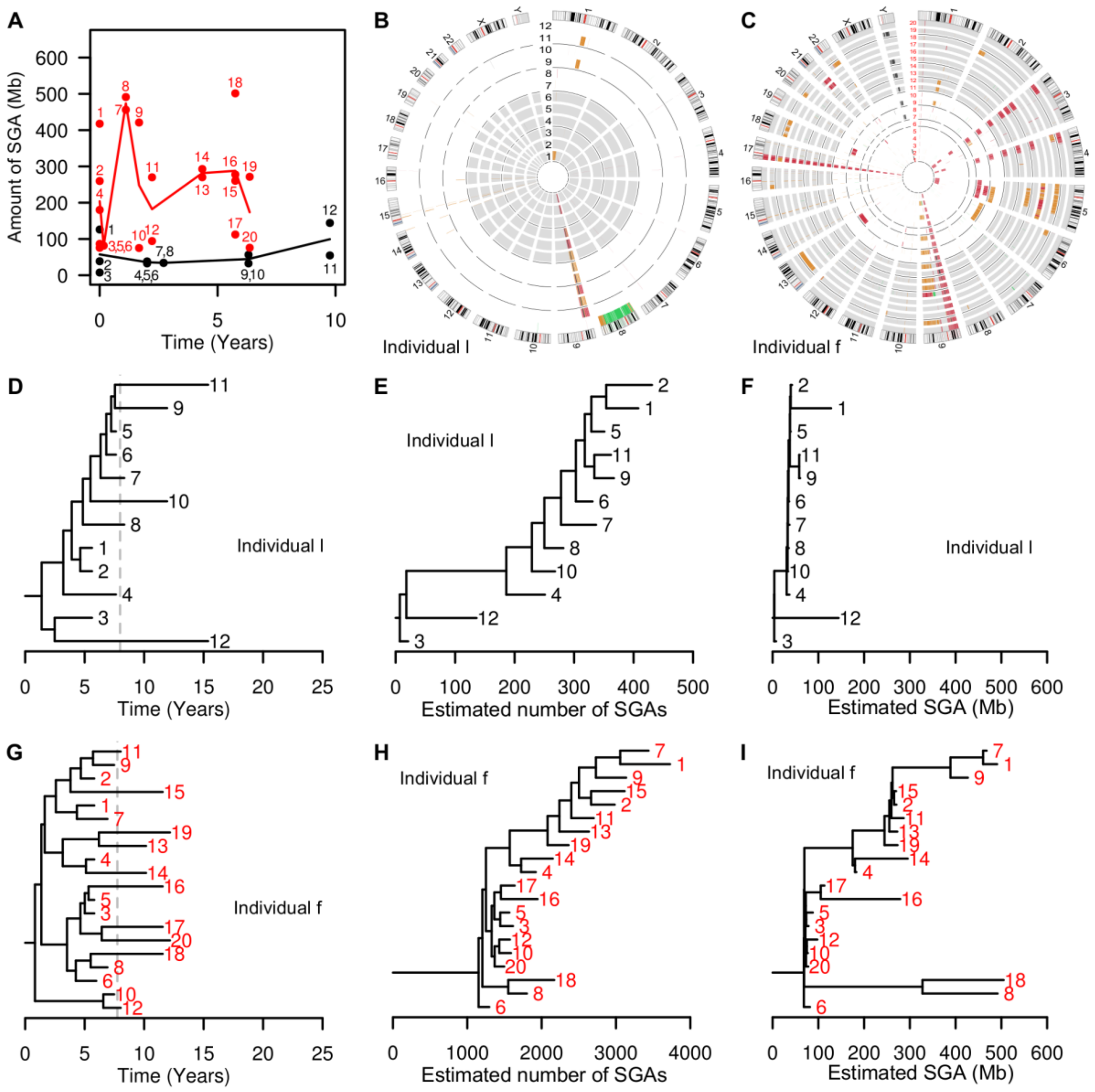


Figure 4.5. Clonal evolution in participants l and f. (Panel A) In individual l (black), the mean amount of SGA was 54 ± 29 Mb over time, whereas in individual f (red) the mean amount of SGA was 242 ± 121 Mb over time. (Panels B and C) Genome-wide view of SGA over time in individuals l and f. (Panel B) During the off-NSAID period in individual l, we detected a whole-chromosome gain of chromosome 8 in biopsy 12 (green band) and some copy-neutral LOH events on chromosome 1 in biopsies 9 and 11 (orange bands). (Panel C) We detected 1,844 ± 573 of SGAs in individual f, who did not progress to EA, but rather opted for esophagectomy for high-grade dysplasia after 6.4 years of follow-up and subsequently died of mesothelioma 11.9 years later. (Panels D, G) Consensus phylogenetic trees estimated by BEAST reveal long-term co-existence of multiple clones. (Panels E, H, F, I) Maximum parsimony trees reveal an underlying progressive evolution of SGA events irrespective of time. Note in individual f that the clade defined by biopsies 1, 7, and 9 seem

the most advanced in progression. Consensus phylogenetic trees generated as indicated in the legend to Figure 4.

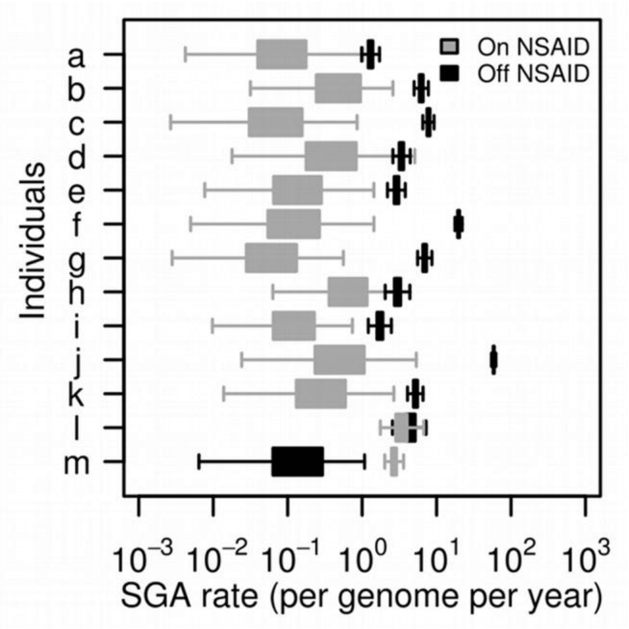
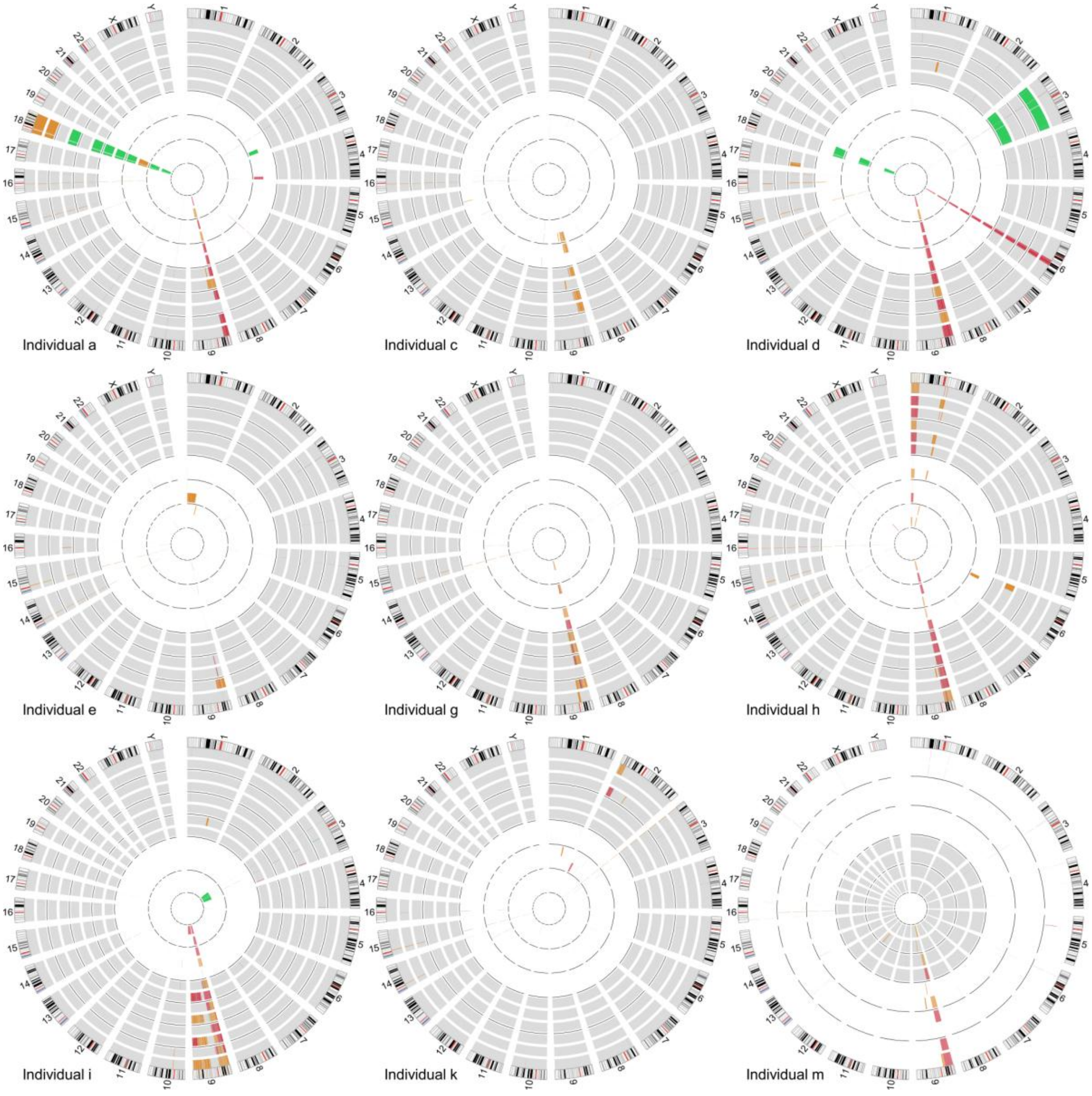
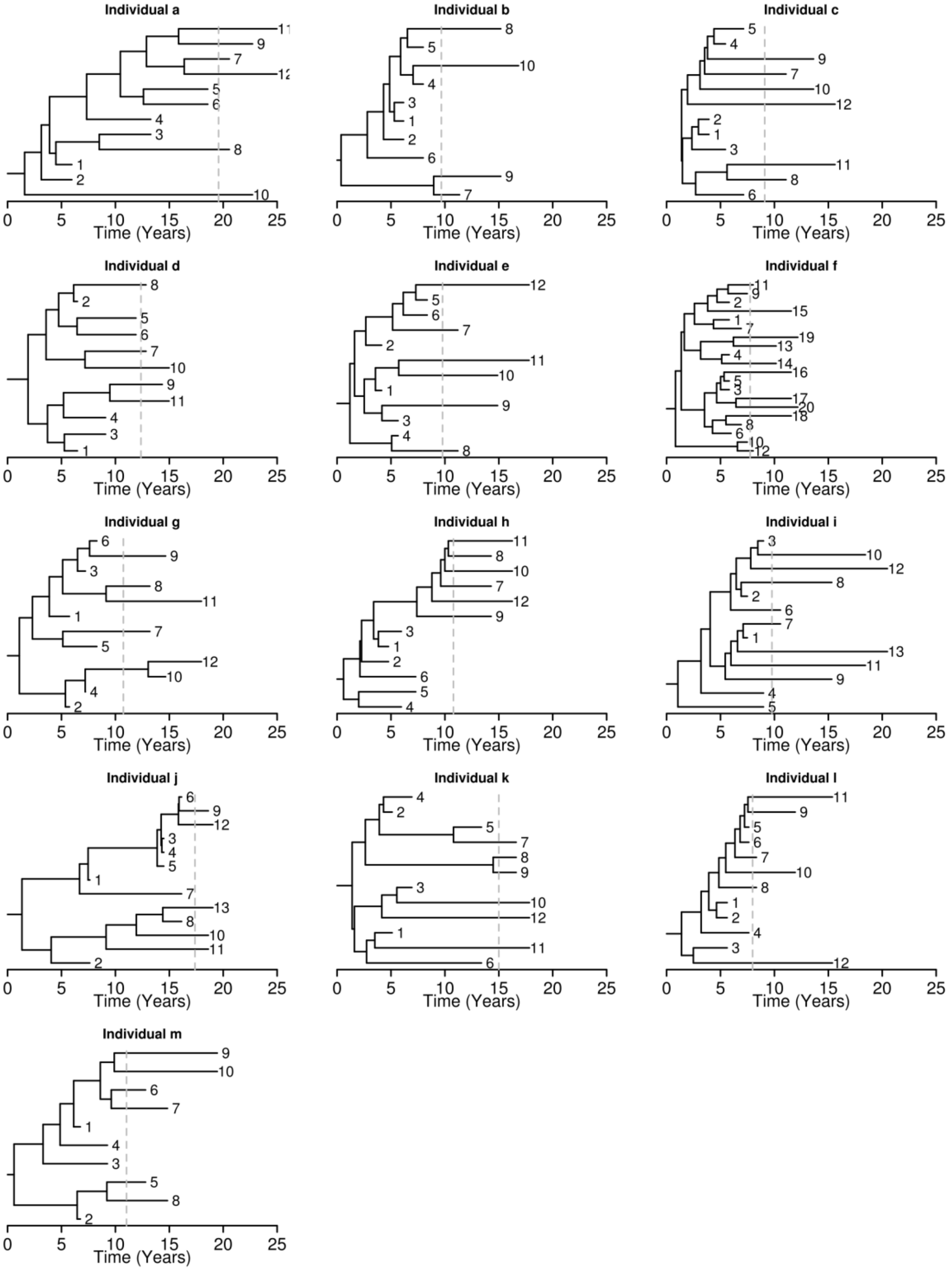


Figure 4.6. BEAST analysis of the SGA patterns across longitudinal biopsies within individuals suggests that NSAID use reduces the SGA rate (number of SGA events per genome, per year). For all individuals (a-m), the mean off-NSAID SGA rate was 7.8 (95% support interval [SI]: 7.1–8.6) and the mean on-NSIAD SGA rate was 0.6 (95% SI: 0.3–1.5). For participants a-k, the mean off-NSAID SGA rate was 8.8 (95% SI: 8.1–9.5,), whereas the mean on-NSAID SGA rate was 0.2 (95% SI: 0.03–1.0). For the two participants l and m that started surveillance on NSAIDs and then went off NSAIDs, there are mixed results. The mean on-NSAID SGA rate for individual l was 3.1 (95% SI: 2.2–4.7) and the mean off-NSAID SGA rate was 4.4 (95% SI: 3.1–5.9). However, for individual m the mean on-NSAID SGA rate was 2.5 (95% SI: 2.1–3.0) and the mean off-NSAID SGA rate was 0.1 (95% SI: 0.01–0.6).

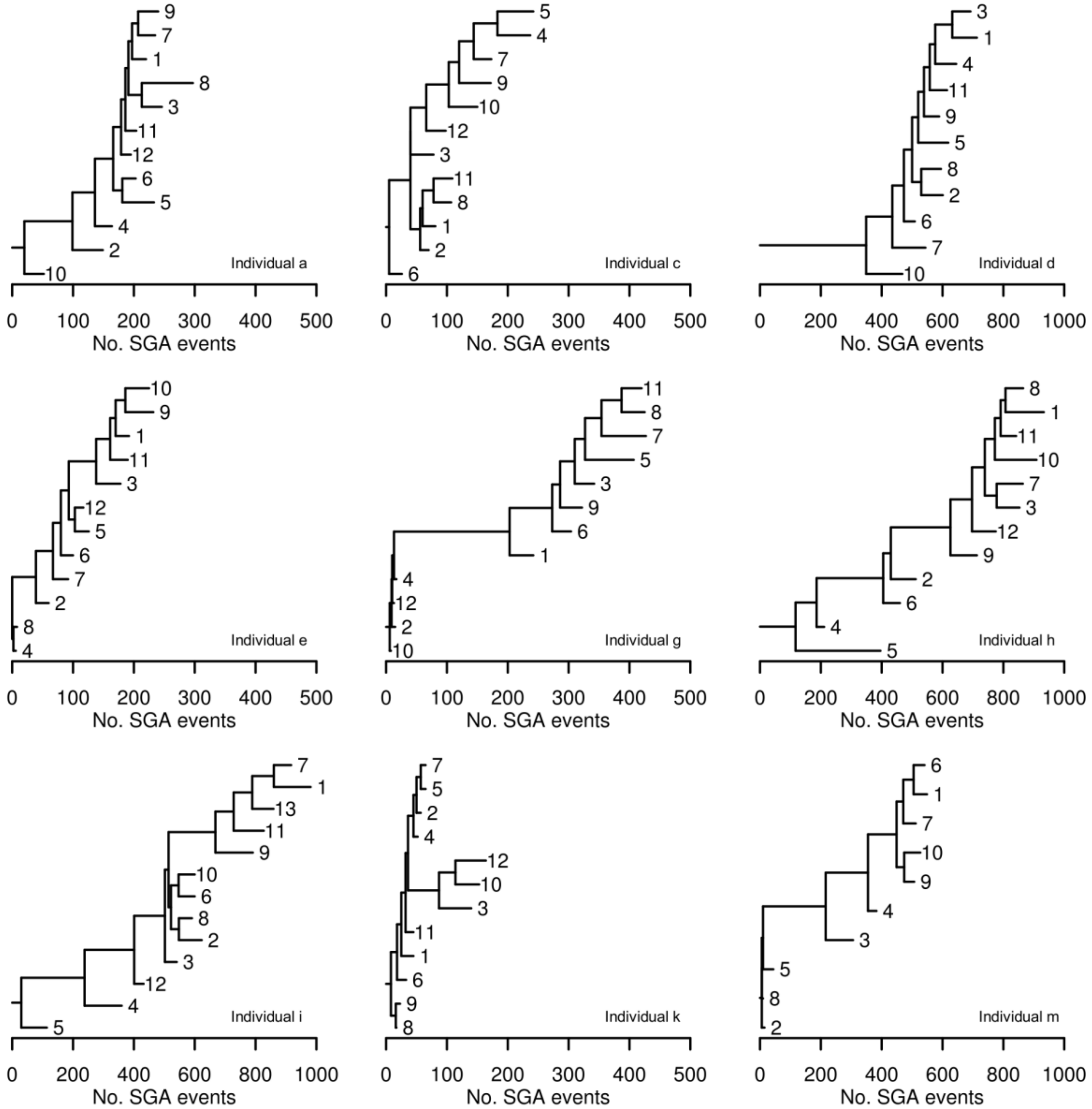
4.7. Supporting information



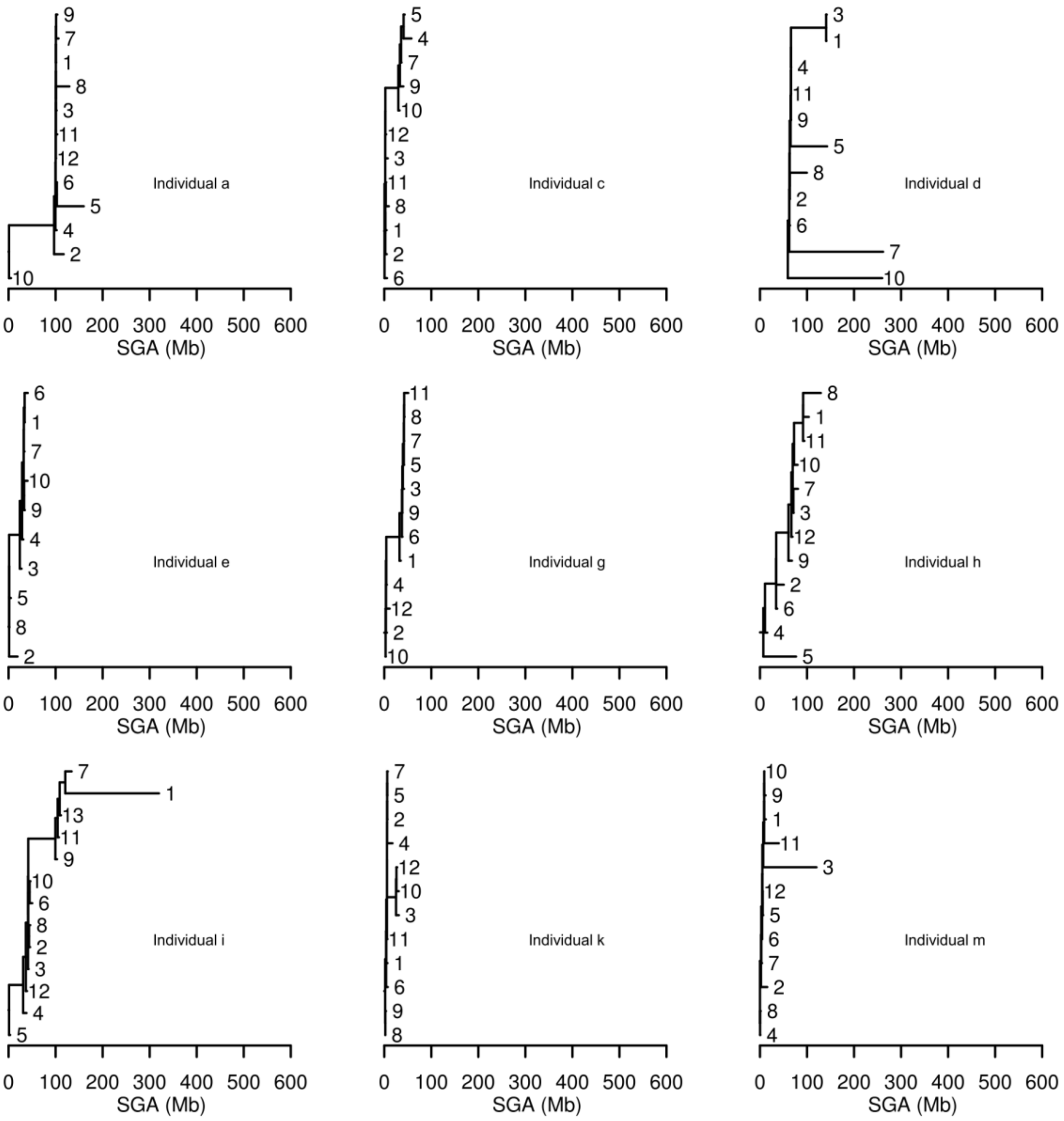
Supplementary Figure 4.7. Circos plots of individuals a,c,d,e,g,h,i,k, and m. Each ring represents whole-genome SGA data from a different biopsy. Thin black line rings separate endoscopies (time points), white background shows time periods off-NSAIDs and gray background shows time periods on-NSAIDs. Within the rings, black segments designate homozygous deletion, red single copy loss, orange copy-neutral LOH, and green shows copy gain.



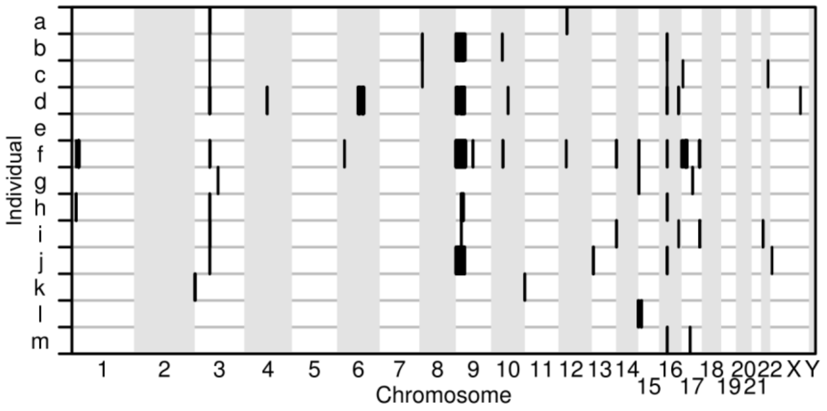
Supplementary Figure 4.8. Estimated trees by BEAST for individuals a-m. Branch lengths are scaled according to time, the tips of the phylogeny are biopsies aligned on the x-axis according to their sampling time, and all internal nodes are estimated coalescence times assuming a logistic population growth model (see Methods). Dashed gray line represents the time point of change in NSAID use. All these trees show long-term co-existence of clones and no evidence of a clonal expansion taking over the Barrett's segment.



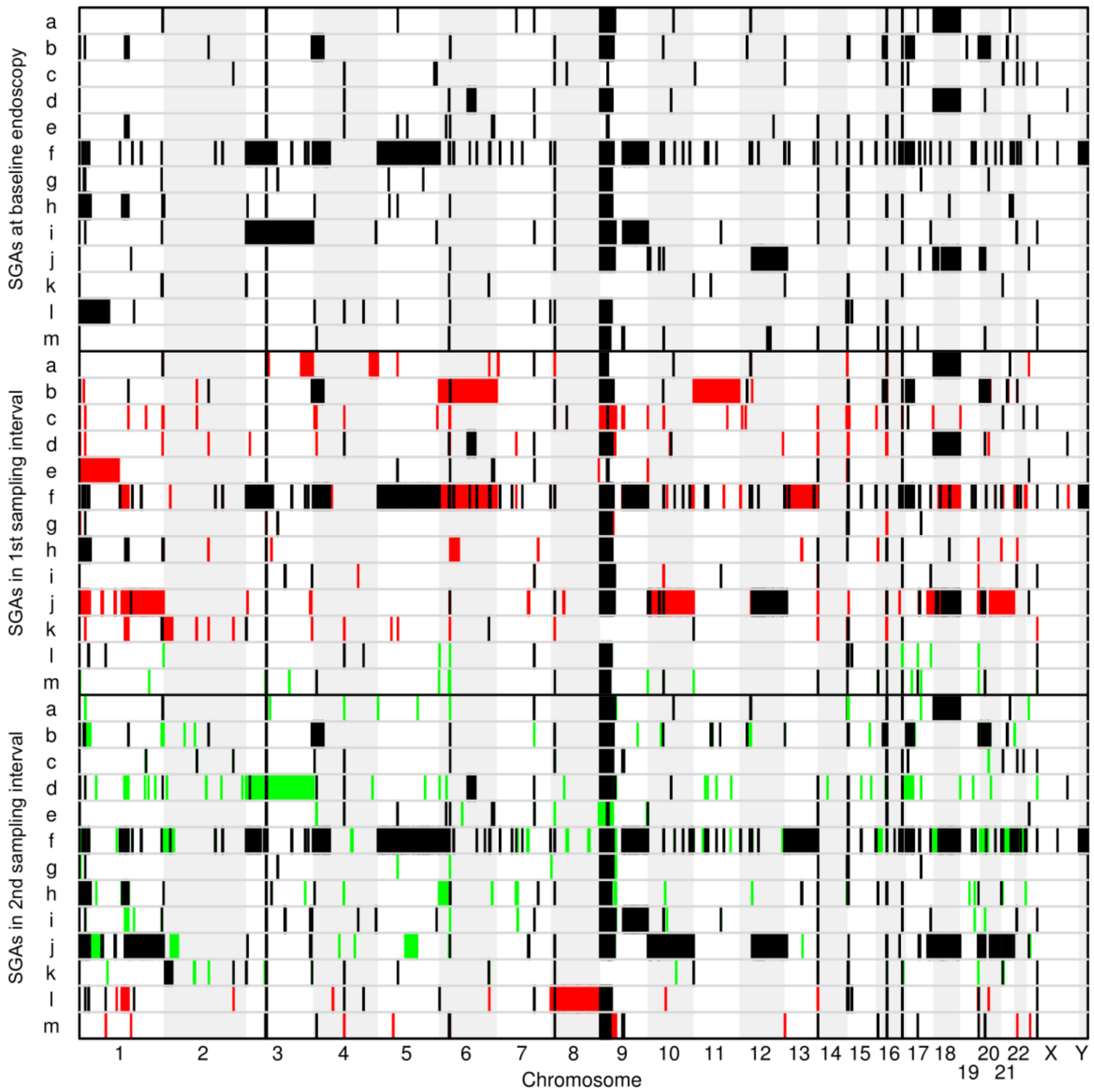
Supplementary Figure 4.9. Estimated trees by PAUP trees where branch lengths represent estimated number of SGA events for individuals a,c,d,e,g,h,i,k and m. The topology of these trees suggest progressive accumulation of SGAs.



Supplementary Figure 4.10. Estimated trees by PAUP trees where branch lengths represent the total amount of SGA (Mb) of the estimated 0->1 SGA events from Supplementary Figure 3 for individuals a,c,d,e,g,h,i,k and m.



Supplementary Figure 4.11. Linear chromosome plot of SGAs that are common across all biopsies within an individual. The most likely regions affected by SGA are fragile sites FHIT and WWOX on chromosomes 3 and 16, respectively, as well as the p-arm of chromosome 9 that often includes the tumor suppressor gene CDKN2A. These are lesions that were present by the time of the first endoscopy and so may have been established with the hypothesized initial expansion of Barrett's epithelium in competition with squamous epithelium.



Supplementary Figure 4.12. Linear chromosome plot of detected somatic genomic abnormalities (SGAs) in biopsies from the baseline endoscopy (top panel), the first sampling period (off-NSAIDs for a-k and on-NSAIDs for l, m; middle panel), and the second sampling period (on-NSAIDs for a-k and off-NSAIDs for l, m; bottom panel).This plot shows the genomic location and size of newly detected SGAs during off-NSAID (red) and on-NSAID (green) periods that are summarized in Figure 3A. Black bars represent SGAs observed in a prior sampling period or at baseline; top panel – black represents SGAs detected in any biopsy from the baseline endoscopy; middle panel – black represents SGAs detected in any biopsy from baseline endoscopy that is also detected in at least one biopsy in the first sampling period; bottom panel – black represents SGAs detected in any biopsy from baseline or first sampling period, or both, that is also detected in at least one biopsy in the second sampling period. In the middle panel, red and green bars represent newly acquired

SGAs that are detected in at least one biopsy in the first sampling period, but not detected at baseline. In the bottom panel, red and green bars represent newly acquired SGAs that are detected in at least one biopsy in the second sampling period, but not detected in any biopsy from baseline or first sampling period.

	To SGA absence (0)	To SGA presence (1)
From SGA absence (0)	$1-\alpha$	α
From SGA presence (1)	0	1

Supplementary Equation S3. Substitution matrix (transition rate matrix) describing the transition probabilities (rates) between SGA absence/presence (0/1) states that was incorporated into BEAST. The probability α is the probability of acquiring SGA in a segment in one division (at the biopsy level), which is estimated by the BEAST software.

$$P(1 \mid 0) = 1 - e^{-\alpha t}$$

$$P(0 \mid 0) = e^{-\alpha t}$$

$$P(1 \mid 1) = 1$$

$$P(0 \mid 1) = 0$$

Supplementary Equation S4. Continuous-time solutions for the probability of all four possible state transitions of a character on a branch of a phylogeny that has branch length of *t* years.

Individual	Sex	Age	L (cm)	A (cm^2)	X	K	N_t	I_b	T_r (days)	T_{init} (years)	BEAST T_{init} upper bound (years)
a	M	44	6	45	3	100,008	58,554	0.14	106	4.96	5.96
b	M	67	8	60	4	11,956	11,519	0.14	103	5.17	6.17
c	M	53	5	37.5	3	47,830	17,220	0.21	73	2.93	3.93
d	M	79	16	120	9	372,496	132,384	0.13	113	5.48	6.48
e	M	53	6	45	4	62,082	28,662	0.21	73	3.13	4.13
f	M	67	12	90	5	31,212	25,213	0.14	103	4.81	5.81
g	M	51	5	37.5	N/A	N/A	N/A	N/A	N/A	4.74*	5.74*
h	M	62	5	37.5	3	33,750	15,000	0.16	93	3.77	4.77
i	M	55	5	37.5	5	44,080	16,840	0.09	161	6.49	7.49
j	M	47	7	52.5	5	90,104	69,671	0.11	133	6.63	7.63
k	M	45	4	30	2	34,147	13,978	0.14	103	4.11	5.11
l	M	53	14	105	3	55,985	33,523	0.14	103	4.62	5.62
m	M	50	5	37.5	3	22,115	11,020	0.1	145	5.75	6.75
Mean		**55.8**	**7.5**	**56.5**	**4**	**72,480**	**36,132**	**0.14**	**109**	**4.81**	**5.81**

Supplementary Table 4.1. Crypt density results. Age is the age of the individual at the first endoscopy. L is the BE segment length measured as the distance in centimeters from the GEJ to the OS anatomical landmarks; A is an estimate of the total area of the BE segment by assuming a circumference of 7.5cm of the esophagus (estimated from [120]); X is the number of levels that were biopsied and had slides from which various numbers of biopsies were evaluated for counting the number of crypts and the number of branching crypts per slide; K is an estimate of the maximum crypt count in BE segments extrapolated from the maximum number of observed crypts in every scored level; N_t is an estimate of the total crypt count in BE segments at baseline endoscopy; I_b is the fraction of crypts that appear to be branching in a sample of crypts; T_r is the estimated crypt doubling time in days (see Equation S1); T_{init} is the estimated time from initiation of the BE segment to baseline endoscopy in years, assuming a logistic growth starting with 1 crypt that grows to a population of N_t crypts at baseline, and assuming a carrying capacity of K crypts for the BE segment (see Equation S2); BEAST T_{init} was bounded to 1 year earlier than the T_{init} to allow some flexibility in the estimate of the exact initiation date during the BEAST MCMC runs. (*) We did not have crypt count information for individual g, so we estimated T_{init} and BEAST T_{init} by taking the average from individuals c,h,i, and m since they had the same segment length (5cm) as individual g.

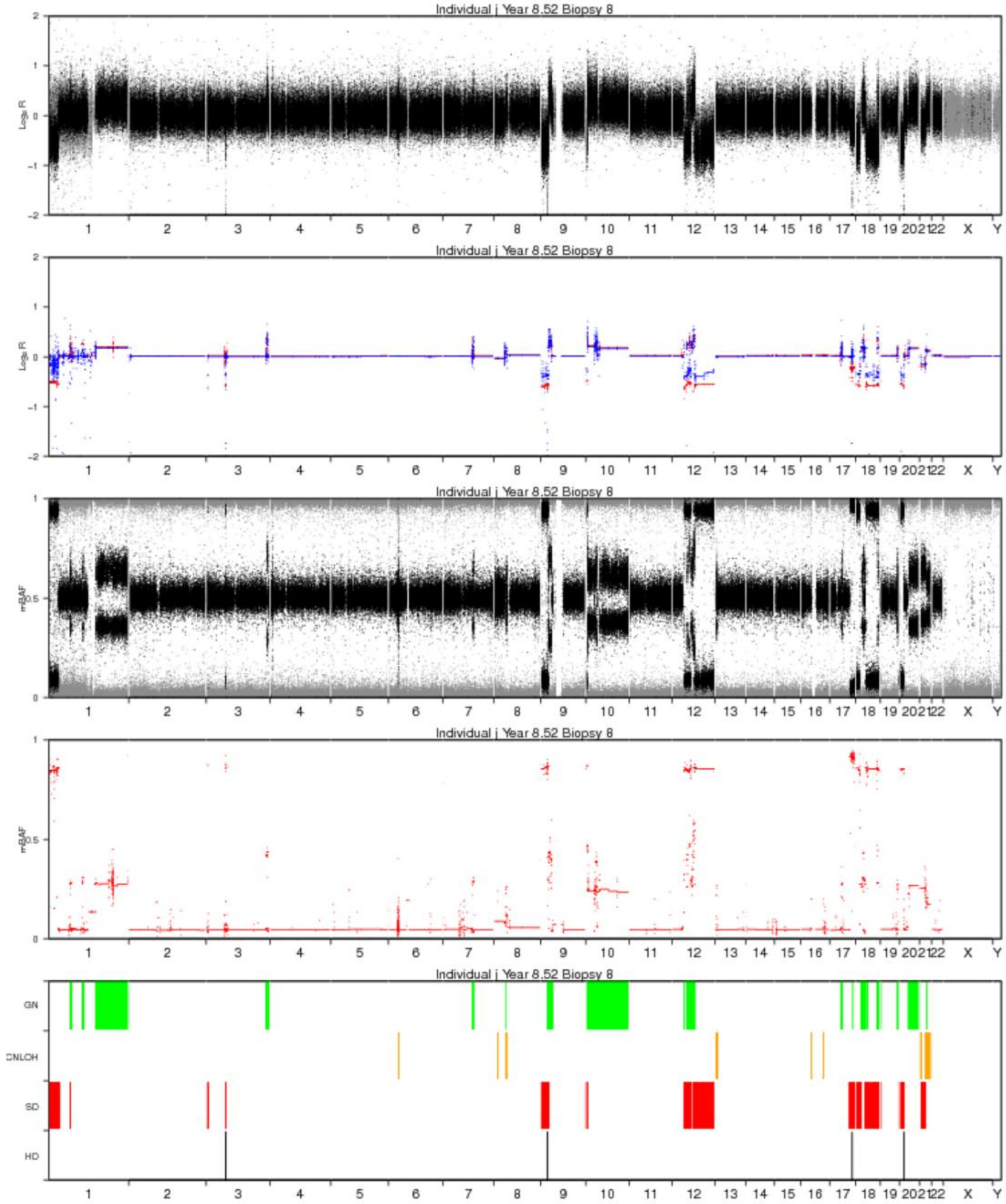
State	Heterozygous SNPs profile Log_2R	Homozygous SNPs profile Log_2R	Allelic imbalance profile mBAF
AB	(-0.2,0.2)	(-0.15,0.2)	(0,0.15)
0	(-3,-0.5)	(-3,-0.5)	(0,0.5)
A	(-3,-0.2)	(-3,-0.15)	(0.15,1)
AA	(-1,0.15)	(-1,0.15)	(0.15,1)
AAA	(1,2)	(1,2)	(0.6,1)
AAB	(0.15,1)	(0.15,1)	(0.15,1)
AABB	(1,3)	(1,3)	(0,0.2)
AAAA	(1,3)	(1,3)	(0.6,1)

Supplementary Table 4.2. Calling thresholds (lower and upper bounds given in parentheses) used for identifying segments with single copy loss, double copy loss, copy gain, and copy neutral LOH.

Individual	**Colless Test**		**Sackin Test**	
	Yule null model	PDA null model	Yule null model	PDA null model
P-values for tree shape imbalance of PAUP* phylogenetic trees				
a	0.004	0.144	0	0.102
b	0.022	0.294	0.004	0.316
c	0.032	0.442	0.012	0.454
d	0	0.068	0	0.112
e	0.006	0.22	0	0.264
f	0.078	0.732	0	0.664
g	0	0	0	0
h	0	0.028	0	0.036
i	0.004	0.228	0	0.16
j	0.058	0.6	0.004	0.548
k	0.018	0.326	0	0.318
l	0	0.022	0	0.022
m	0	0.07	0	0.062
P-values for tree shape imbalance of BEAST phylogenetic trees				

a	0.03	0.442	0.006	0.376
b	0.124	0.552	0.018	0.498
c	0.772	0.968	0.13	0.93
d	0.95	0.996	0.258	0.998
e	0.492	0.918	0.08	0.868
f	0.664	0.992	0.154	0.978
g	0.448	0.882	0.096	0.904
h	0.016	0.26	0.002	0.3
i	0.034	0.49	0	0.444
j	0.176	0.714	0.042	0.79
k	0.98	1	0.294	1
l	0.01	0.27	0.002	0.316
m	0.158	0.696	0.044	0.658

Supplementary Table 4.3. Tree shape imbalance statistics for individuals a–k estimated for PAUP* and BEAST trees calculated from generating 500 random trees having the same number of taxa under Yule and PDA null models [121] using the R package "apTreeshape" [122]. Significant p-values reject the null hypothesis that the observed tree shape is as balanced as the 500 random tree shapes, where the test statistics are the Colless and Sackin formulas [121] for calculating tree shape imbalance.



Supplementary Figure 4.13. Raw data segmentation and SGA calls for all 161 biopsies of individuals a–m. Every page shows an individual biopsy and has 5 panels (top to bottom): first panel, raw Log_2R ratio between biopsy and leukocyte control, where gray SNPs are homozygous SNPs and black SNPs are heterozygous SNPs; second panel, GLAD segmentation of homozygous SNPs (blue line) and heterozygous SNPs (red line); third panel, raw mBAF (reflected and scaled B Allele Frequency of the BE sample) where homozygous SNPs are shown in gray, and heterozygous SNPs are shown in black; fourth panel, GLAD segmentation of the mBAF data of heterozygous SNPs, which are informative for allelic imbalance; fifth panel, final SGA calls for chromosomal regions: GN (copy gain, green), CNLOH (copy neutral LOH, orange), SD (single deletion, or single copy number loss, red), HD (homozygous deletion, or double copy number loss, black). Only one example biopsy is displayed here, which is biopsy #8 from individual j, which shows massive SGA.

4.7.1. Cell organization into crypts in Barrett's Esophagus and Analytical modeling of crypt production in Barrett's esophagus (Supplementary Text S1)

To understand how evolution at the genome level unfolds, we need to first consider how cells are spatially organized in the BE tissue. The BE epithelium is a single layer of specialized intestinal metaplasia that typically covers 38 cm^2 and contains between 10,000-400,000 crypts (Supplementary Table 4.1). Crypts hold a reservoir of stem cells at their base that are self-renewing cells capable of generating and propagating genomic alterations over long timespans. Cell proliferation and differentiation occurs at the base of the crypts producing a flux of differentiated cells that move up the crypt and slough off into the lumen. Because of this constant shedding of cells, cell fitness in a crypt-structured epithelium is a complex combination of stem cell self-renewal, survival, and lateral spread, taking over neighboring crypts. Acquired SGAs that boost any of these three cell phenotypes will persist and increase in frequency in the BE cell population. The organization of cells into crypts in itself is an evolved mechanism to protect against cancer since acquired SGAs in differentiated cells can be lost by cells sloughing off and only acquired SGAs in stem cells can persist and increase in frequency over time; and also, in a strictly single-layer columnar epithelium, lateral invasion of mutant cells into neighboring crypts is physically difficult since invading from the lumen side requires going against the flux of cells [17]. A subset of SGAs were detected in all biopsies of an individual over decades (Supplementary Figure 4.11) which is evidence that these SGAs must be acquired in long-living self-renewing (stem) cells that have not differentiated, been sloughed into the lumen, and have survived toxic microenvironmental exposures. Other SGAs had various lifespans in the segment (Figures Figure 4.2, Figure 4.4B,C, Figure 4.5B,C, Supplementary Figure 4.12) but only a fraction of them were turning over, either appearing or regressing (Figure 4.3). In summary, the spatial organization of cells into a single-layer crypt-structured BE epithelium constrains the evolutionary dynamics and spatial dispersal of acquired SGAs over time.

$$T_r = \frac{T_b \log 2}{\log(1 + I_b)}$$

Supplementary Equation S1. Estimating the crypt doubling time where the duration of branching (T_b) is set to 20 days; see Appendix I from [97]. I_b is the fraction of crypts that appear to be branching in standard pathology slides of BE biopsies.

The duration of the crypt replication cycle T_r can also be defined as a crypt doubling time, in both definitions T_r is the time it takes, in days, for a single crypt to replicate and produce two daughter crypts. This equation assumes exponential growth, where the number of crypts at time t is [123]:

$$N(t) = N(0)e^{\frac{\log 2}{t_r}t}$$

Often, population growth is expressed using a growth rate parameter, or Malthusian growth parameter, m, which is:

$$m = \frac{\log 2}{t_r}$$

Although simple mathematically, exponential growth is not realistic for biological systems, where when populations grow in size and density, the competition for space and resources increases, which causes growth to slow down. Such behavior can be modeled with a logistic growth equation, which is one type of sigmoid functions, which introduces the concept of carrying capacity, or maximal population size. As the population size reaches its carrying capacity, its growth rate decreases.

In Barrett esophagus, the date of the most recent common ancestor (MRCA) of all biopsies (or time from initiation to baseline endoscopy) can be identified if logistic growth is assumed. For the logistic growth, the Barrett segment is initiated with a single crypt, i.e. at time T_{init}, the crypt population size is N_0=1. At time t of baseline endoscopy, the crypt population size is N_t, and for the duration of evolution, the maximal crypt population size is K. The following logistic equation describes the duration of the initiation phase T_{init} which is in units of days, since T_r is also in units of days:

$$T_{init} = \frac{T_r}{\log 2} \log \frac{N_t(K - N_0)}{(K - N_t)N_0}$$

Supplementary Equation S2. Estimating the initiation phase duration, where N_0=1 is the starting population size that is fixed to 1 crypt. *K* is the estimated maximum number of crypts ("carrying capacity") in a BE segment , N_t is the estimated total number of crypts in a BE segment, and T_r is the crypt doubling time defined in Supplementary Equation S1.

Where Ib is the fraction of branching crypts, calculated by taking the average number of branching crypts for each level and dividing by the average number of crypts for that level (for each level, 1-4 pathology slides were counted), choosing the level with maximum number of branching crypts. L is the segment length (cm), or the distance between the lower esophageal sphincter (LES) and ora serrate (OS) landmarks. A is the surface area, calculated by multiplying L and the circumference of the esophagus, which is fixed at 7.5 cm (approximately the diameter of a US quarter coin). K is the carrying capacity, calculated by taking the maximum number of crypts detected in any one of the slides within a level, and summing those counts across levels to get the maximum possible crypt count per segment. N_t is the average crypt count per level, calculated by taking the crypt count for each slide and dividing by the number of slides, and extrapolating to the level surface area. X is the number of levels that were scored by pathology out of L possible levels.

Comparison of % branching crypts in normal and diseased epithelium

I_b is the observed fraction of crypts that appear to be budding or branching in a sample of crypts, isolated from a crypt-structured epithelium (Barrett's epithelium, small intestinal or colonic epithelium). Often, I_b is measured by incubating the tissue in EDTA and peeling off the epithelium from stroma, after which, crypts could be scored as budding or branching by visual examination under a dissecting microscope. Alternatively, I_b can be estimated from H&E stained histopathology slides, however, only crypts that show the majority of the crypt lumen, ideally both the crypt bottom and the crypt mouth, can be safely counted as budding or branching. For simplicity, the term "branching" would mean crypts that are either budding or branching, since both appearances are part of the branching process, therefore I_b counts both budding and branching crypts.

My estimates of I_b from Srivastava and Odze's pathology slide measurements ranged 9%-21% (mean 14%) and fall within the ranges of two previously reported crypt branching rate estimates in the literature. Cheng et al. reported an average of 0.44%, 30.4%, 15.1%, and

13.2% branching crypts in 11 normal adults, and 4 Ulcerative colitis, 4 Crohn's disease, 4 Multiple polyposis patients [97]. The average ages and age ranges of the four groups were 69 (48-83), 47(31-56), 33 (27-45), and 42 (34-56) [97]. Cummins et al. reported an average of 7.8%, 15%, 4.9%, 1.7% branching crypts in small intestine in 3 neonates, 16 infants, 14 children, and 39 adults [124]. The average ages and age ranges of the four groups were 2.4 weeks (0.9-4 weeks), 0.7 years (0.3-1.7 years), 7.9 years (2.4-16.2 years), and 46 years (20-80 years) [124]. Therefore, it is likely that in hyperproliferative Barrett tissues crypts attain a rate of branching similar during pregnancy and infancy, surpasses the branching rate in adult normal intestinal epithelium, and compares to rate of branching in diseased colonic epithelium.

4.7.2. Detecting mutation rate change in Barrett's esophagus after treatment with NSAIDs: computational power analysis

Authors: Rumen Kostadinov, Mary Kuhner, Carlo Maley

4.7.2.1. Introduction and Methods

I performed *in-silico* analysis in 2007 prior to conducting the large study in this Chapter to calculate the power to detect mutation rate changes, which was critical for selecting the number of time points and the number of biopsies per time point to assay per individual patient, given the high cost of whole-genome SNP assays. I modified SerialSimCoal [63] to simulate an evolving tumor cell population with the following parameters:

24 samples over 10 years
4 samples every 2 years (=6 time points)
300 cell generations per year
8,000 neutral loci (no natural selection)
1.1, 2, 5, 10 and 100-fold decrease in mutation rate after year 5
AR – ancient mutation rate (off NSAIDs) (5 year duration)
RR – recent mutation rate (on NSAIDs) (5 year duration)
Effective population size, N_e = 100, 1000, 10000
Constant population size over time
Population-scaled mutation rate (theta) θ = 0.000082, 0.00082, 0.0082

Table 4.4. Parameter table for the *in-silico* power analysis.

I used BEAST [62,107] to estimate population sizes and mutation rates before and after year 5, given 24 samples. I adapted SimCoal and BEAST to use sequences of 0s and 1s representing absence and presence of LOH, and used a novel two-state LOH substitution model. I adapted the BEAST software for coalescent-based backward estimation of population genetic parameters given SGA data. I made three major changes to BEAST: addition of a 0/1 binary evolutionary model, addition of a sequence error, and addition of an extra MRCA-LUCA branch.

First, I included a 0/1 binary evolutionary model that instead of having four DNA states A,C,G, and T, which are modeled with DNA evolutionary models such as Jukes-Cantor or F84, has only 2 possible states, 0 and 1, and allows 0->1 and 1->0 transitions. If the probability of 0->1 transition is alpha and the probability of 1->0 transition is beta for one step of the Markov chain then a continuous approximation can be derived for the state of the chain after t steps, where t can be a decimal (Dr. Joseph Felsenstein, personal communication). I implemented this 0/1 model into BEAST allowing the user to specify alpha and beta as parameters. Notably, if we would like to make an irreversible model by disallowing 1->0 transition, we simply set the 1->0 probability beta to 0. The 0/1 model, as any other evolutionary model, simply calculates the likelihood of starting at a given state at a parent node in a phylogeny and ending at any other state at a child node in a phylogeny given a certain branch length or distance between the parent and child nodes.

Second, I implemented a sequence error in BEAST. An observed state in a sequence could be incorrect prone for example due to instrument technical errors. Typically in DNA evolutionary models a sequence error is modeled by replacing the likelihoods of A, C, G, T at a site (terminal node, or tip) that is called "A" by the assay from 1, 0, 0, 0 to 1-e, e, e, e, where e is the sequence error [125].

I adapted SerialSimCoal software [63] for coalescent-based forward simulation of SGA data given population genetic parameters. SerialSimCoal is a forward simulation of population dynamics and various evolutionary models can be used to simulate serial samples given a specified demographic history. I made changes in SerialSimCoal to produce 0/1 data where a change from 0->1 is irreversible, and where the ancestral state of the population is always 0.

The modified versions of BEAST and SerialSimCoal are available upon request.

4.7.2.2. Results and Discussion

For 91% (41/45) of the parameter combinations, we had >80% power to detect a 1.1- to 100- fold decrease in mutation rate (paired Wilcoxon test, $p<0.05$) (Table 4.5). As the fold difference in mutation rate increased between ancient and recent periods, the number of parameter sweeps that accurately estimated the simulated mutation rates increased (third and last column of Table 4.5). Overall, BEAST recovered mutation rates accurately from simulated data, which verified that the implementation of the novel 0/1 evolutionary model was accurate and error-free.

True Ne	Theta	Fold Diff.	Ne 95% CI lower	Ne 95% CI upper	True AR	AR 95% CI lower	AR 95% CI upper	True RR	RR 95% CI lower	RR 95% CI upper	No. sweeps cover. Ne	No. sweeps cover. AR	No. sweeps cover. RR	No. sweeps AR>RR p<.05 n=100
100	8.25E-05	1.1x	45	277	4.3E-07	2.2E-07	9.1E-07	3.9E-07	1.9E-07	8.7E-07	93	97	93	91
100	8.25E-05	2x	48	297	5.8E-07	2.9E-07	1.0E-06	2.9E-07	1.3E-07	7.2E-07	93	93	96	90
100	8.25E-05	5x	49	271	9.2E-07	5.1E-07	1.5E-06	1.8E-07	7.1E-08	5.6E-07	93	93	95	99
100	8.25E-05	10x	44	237	1.3E-06	8.4E-07	2.1E-06	1.3E-07	4.3E-08	4.6E-07	96	94	97	100
100	8.25E-05	100x	52	208	4.1E-06	3.1E-06	5.3E-06	4.1E-08	1.3E-08	3.2E-07	95	89	92	100
100	8.25E-04	1.1x	64	183	4.3E-06	3.3E-06	5.6E-06	3.9E-06	2.9E-06	5.0E-06	94	98	93	83
100	8.25E-04	2x	62	178	5.8E-06	4.7E-06	7.4E-06	2.9E-06	2.1E-06	3.9E-06	91	96	97	100
100	8.25E-04	5x	63	204	9.2E-06	7.6E-06	1.1E-05	1.8E-06	1.2E-06	2.7E-06	91	97	97	100
100	8.25E-04	10x	62	178	1.3E-05	1.1E-05	1.5E-05	1.3E-06	8.1E-07	2.1E-06	93	94	96	100
100	8.25E-04	100x	205	1033	4.1E-05	3.4E-05	4.2E-05	4.1E-07	1.8E-07	8.5E-07	89	90	94	100
100	8.25E-03	1.1x	322	1544	4.3E-05	3.3E-05	4.3E-05	3.9E-05	3.6E-05	4.4E-05	81	87	86	97
100	8.25E-03	2x	307	1615	5.8E-05	4.4E-05	5.5E-05	2.9E-05	2.7E-05	3.3E-05	83	86	86	100
100	8.25E-03	5x	584	4449	9.2E-05	5.6E-05	8.1E-05	1.8E-05	1.7E-05	2.3E-05	79	77	82	98
100	8.25E-03	10x	1064	7055	1.3E-04	7.6E-05	1.1E-04	1.3E-05	1.2E-05	1.6E-05	78	75	91	100
100	8.25E-03	100x	2814	25807	4.1E-04	5.4E-05	2.1E-04	4.1E-06	2.9E-06	5.4E-06	46	26	94	99
1000	8.25E-05	1.1x	407	18016	4.3E-08	8.9E-09	1.6E-07	3.9E-08	9.7E-09	1.7E-07	96	98	96	82
1000	8.25E-05	2x	401	12370	5.8E-08	1.8E-08	2.0E-07	2.9E-08	8.0E-09	1.7E-07	95	96	95	82
1000	8.25E-05	5x	467	7413	9.2E-08	3.5E-08	2.5E-07	1.8E-08	6.0E-09	1.6E-07	96	92	93	95
1000	8.25E-05	10x	448	3675	1.3E-07	6.0E-08	3.3E-07	1.3E-08	3.3E-09	1.4E-07	97	94	96	98
1000	8.25E-05	100x	564	2095	4.1E-07	2.4E-07	7.0E-07	4.1E-09	2.6E-09	1.3E-07	98	95	77	100
1000	8.25E-04	1.1x	615	2082	4.3E-07	2.5E-07	7.0E-07	3.9E-07	2.1E-07	6.8E-07	97	93	93	88
1000	8.25E-04	2x	653	2079	5.8E-07	3.6E-07	8.8E-07	2.9E-07	1.5E-07	5.6E-07	92	92	96	98
1000	8.25E-04	5x	614	1867	9.2E-07	6.3E-07	1.3E-06	1.8E-07	8.9E-08	4.3E-07	94	95	93	100
1000	8.25E-04	10x	635	1834	1.3E-06	9.2E-07	1.8E-06	1.3E-07	5.4E-08	3.5E-07	97	93	98	100
1000	8.25E-04	100x	695	1832	4.1E-06	3.4E-06	5.0E-06	4.1E-08	1.5E-08	2.1E-07	88	96	95	100
1000	8.25E-03	1.1x	699	1729	4.3E-06	3.6E-06	5.3E-06	3.9E-06	3.2E-06	4.8E-06	91	94	98	91
1000	8.25E-03	2x	726	1778	5.8E-06	4.9E-06	6.8E-06	2.9E-06	2.3E-06	3.7E-06	93	98	95	100
1000	8.25E-03	5x	690	1685	9.2E-06	8.0E-06	1.0E-05	1.8E-06	1.4E-06	2.5E-06	97	95	98	100
1000	8.25E-03	10x	703	1714	1.3E-05	1.2E-05	1.5E-05	1.3E-06	9.1E-07	1.9E-06	96	93	97	100
1000	8.25E-03	100x	695	2108	4.1E-05	3.7E-05	4.4E-05	4.1E-07	2.3E-07	7.5E-07	95	95	96	100
10000	8.25E-05	1.1x	1588	67909	4.3E-09	8.7E-10	6.9E-08	3.9E-09	1.8E-09	9.0E-08	96	99	95	80
10000	8.25E-05	2x	1868	74827	5.8E-09	1.1E-09	6.6E-08	2.9E-09	1.4E-09	8.4E-08	97	99	87	75
10000	8.25E-05	5x	2646	74952	9.2E-09	1.8E-09	6.1E-08	1.8E-09	1.1E-09	7.1E-08	98	98	85	75
10000	8.25E-05	10x	3261	73887	1.3E-08	2.5E-09	6.7E-08	1.3E-09	9.5E-10	7.1E-08	98	96	85	73
10000	8.25E-05	100x	4738	65189	4.1E-08	1.0E-08	1.1E-07	4.1E-10	1.6E-09	8.8E-08	93	92	10	81
10000	8.25E-04	1.1x	5175	63593	4.3E-08	1.1E-08	1.1E-07	3.9E-08	1.2E-08	1.4E-07	94	93	95	82
10000	8.25E-04	2x	4659	54874	5.8E-08	1.6E-08	1.3E-07	2.9E-08	7.0E-09	1.3E-07	97	98	95	83
10000	8.25E-04	5x	5350	51011	9.2E-08	2.7E-08	1.8E-07	1.8E-08	5.4E-09	1.3E-07	92	94	92	92
10000	8.25E-04	10x	5731	41713	1.3E-07	5.2E-08	2.5E-07	1.3E-08	4.5E-09	1.4E-07	92	90	91	90
10000	8.25E-04	100x	6425	26243	4.1E-07	2.2E-07	6.0E-07	4.1E-09	3.1E-09	1.6E-07	89	92	78	99
10000	8.25E-03	1.1x	6557	28239	4.3E-07	2.1E-07	6.4E-07	3.9E-07	2.2E-07	6.7E-07	96	94	92	85
10000	8.25E-03	2x	6750	24801	5.8E-07	3.3E-07	8.4E-07	2.9E-07	1.3E-07	5.5E-07	96	92	97	96
10000	8.25E-03	5x	6769	21222	9.2E-07	5.7E-07	1.2E-06	1.8E-07	7.5E-08	4.6E-07	95	95	93	100
10000	8.25E-03	10x	6771	19895	1.3E-06	8.7E-07	1.6E-06	1.3E-07	4.5E-08	4.2E-07	94	95	93	100
10000	8.25E-03	100x	7057	17568	4.1E-06	3.4E-06	4.8E-06	4.1E-08	1.3E-08	3.3E-07	94	94	94	100

Table 4.5 BEAST inference of mutation rates on and off NSAIDs and of constant population size from samples generated from a SerialSimCoal forward simulation using a fixed set of mutation rates on and off NSAIDs and a fixed constant population size.

4.7.2.3. Conclusion

This power analysis helped in getting grant funding and in study design of the larger experiment that tested the hypothesis that an association between mutation rate change and NSAIDs use exists. This power analysis was presented at the AACR Frontiers in Cancer Prevention conference in Philadelphia in 2007.

4.8. Chapter conclusion

In summary, our results suggest that in most individuals in long-term endoscopic surveillance with BE, clonal evolution within the BE segment has occurred prior to baseline detection in the clinic, but from baseline detection to more than a decade of follow-up the population remains in relative evolutionary stasis at the genome level. Only rarely is evolutionary stasis punctuated by a massive burst of SGA that may lead to progression to cancer. The current picture supports an evolutionary scenario in which mutation-selection balance of spontaneous SGAs and purifying selection against deleterious effects maintain a relative stasis of the genome until a chance occurrence of adaptive (for the clone) SGA, perhaps with epistatic effects, or changes in the selective environment, results in a new proliferating clone that may progress to EA. In addition to the observation of evolutionary stasis, NSAID use in Barrett's esophagus is associated with a reduction in the rate of acquisition of SGA suggesting that the pathway whereby NSAIDs exert their protective effect involves the reduction in number of SGAs or the inhibition of spread of SGA-containing clones. Finally, detection of evolutionary stasis might be used in the clinic to reduce overdiagnosis and unwarranted treatment and detection of massive bursts of SGA might be used to better identify patients needing more aggressive surveillance and therapy.

Chapter 5. Pilot *in vitro* experiment evaluating the genotoxic effect of deoxycholic acid on the evolutionary dynamics of SGA

Author: Rumen Kostadinov

5.1. Introduction

Studying the evolutionary dynamics of SGA *in vivo* in humans has many limitations due to the infeasibility for performing a fully controlled experiment. Ideally, we aim to study the evolution of cell lineages carrying acquired SGA under the presence or absence of genotoxic constituents of duodenal gastroesophageal reflux. However, *in vivo*, administering genotoxic compounds in a controlled experiment is unethical and unfeasible. Therefore, we designed an experimental evolution study to evaluate the *in vitro* SGA evolutionary dynamics in a Barrett's esophagus cell line, where we controlled the dose and duration of genotoxic stress and observed the acquisition of SGA in evolving cell lineages over time. Ultimately, modulating the dose and duration of genotoxic stress corresponds to modulating the mutation rate, or the rate of acquisition of SGA, which is a key population-genetic parameter providing a quantity of the raw material that allows for evolution by natural selection to occur. Moreover, we aimed to develop computational and statistical methods for estimating SGA rate in order to test whether the inferred SGA rate from observed patterns of SGA *in vitro* correlates positively to the level of genotoxic exposure to which we subject the cell line. In summary, the experimental evolution study allowed us to evaluate the relationship between the level of DCA genotoxicity, the level of detected SGAs, and the evolution of SGAs that we define as the change in the frequency of SGA over an approximately similar number of cell generations under presence of various levels of DCA or absence of DCA exposure.

Jenkins et al. showed that the bile acid deoxycholic acid (DCA) is genotoxic and may be associated with the acquisition of SGA and carcinogenesis in BE [10,126]. Therefore, we selected DCA as the source of mutational input to the cell lines. Jenkins et al. tested various concentrations of DCA, ranging from 0-400µM in an approximately fixed concentration of esophageal cells (10^5 cells/mL) (See Fig. 1 in Ref.[126]). I also tested the effect of various concentrations of DCA on relative cell viability, since we aimed to achieve low and high mutation rates that both retain cell viability so that we can assay the cells' genomic state over time.

5.2. Methods

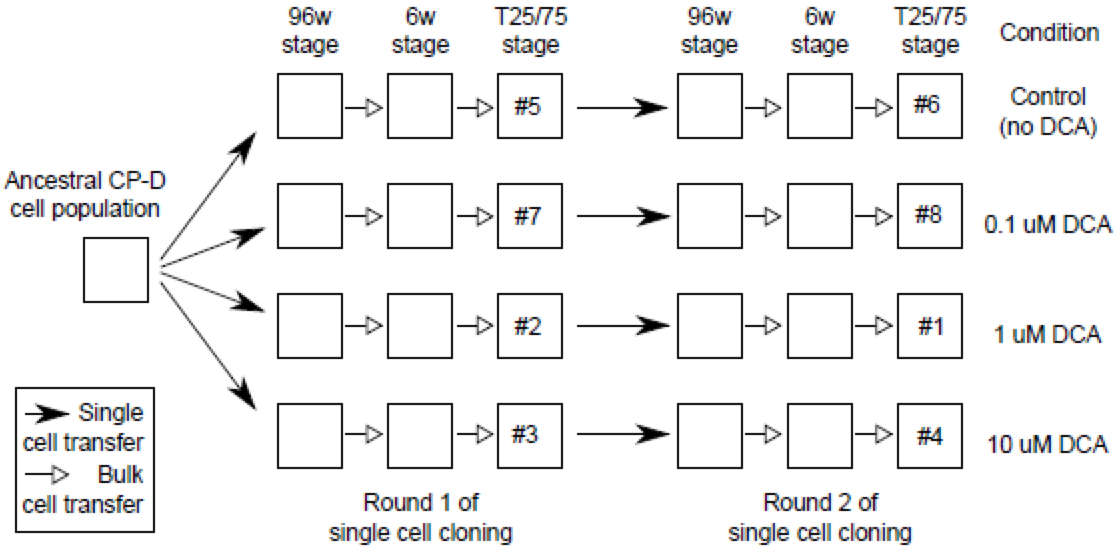


Figure 5.1. Study design of single cell experimental evolution. Single cells from an ancestral cell population from a CP-D Barrett's cell line were grown in 96 well plates, where cells were plated at a concentration of 1/3 of cells per well. Cells were passaged into larger 6-well plates and T25 and T75 plates as they grew in number. After two rounds of single cell cloning, DNA was extracted from 8 samples (samples #1-#8) and DNA was evaluated with Illumina

550K arrays at the Penn Hospital of the University of Pennsylvania (HUP) genomics facility. The two rounds of single cell cloning ensure that any difference in genomic state between samples #5 and #6, #7 and #8, #2 and #1, and #3 and #4 occurred during the course of the experiment (under the various deoxycholic acid (DCA) concentrations).

BE cell line ID	CP-D
BE cell line name	CP-18821
Method of Initiation of Primary Cultures	Enzymatic
Biopsy extracted from region of:	High-grade dysplasia
Serum-Free adapted, Keratinocyte serum-free media (Invitrogen catalog no. 17005-042)	Yes
Doubling Time (hrs)	32.26
DNA content	2.43 (4.56 subpopn)
% 4N DNA	28.5
Tetraploid (%)	15.8
17p LOH	Yes (in vitro) No (in vivo)
P53 mutation	Frameshift 302
P53 mutation	Not detected in vivo
9p LOH	Homozygous deletion (4N, 8N)
CDKN2/p16	Deletion of C5.1 STS marker (4N, 8N)
5q LOH	No
Fluorescently tagged populations	GFP, DsRed

Table 5.1. CP-D Barrett's esophagus cell line description

I used a Barrett's esophagus cell line named "CP-D" or "CP-18821", which was described in Palanca-Wessels et al. (Table 5.1) (M. C. Palanca-Wessels et al. 1998; M. C. A. Palanca-Wessels et al. 2003).

CP-D cells were cultured in KSFM media (Invitrogen) supplemented with 25mg of bovine pituitary extract, 2.5ug recombinant epidermal growth factor (Invitrogen), 50 units of Penicillin, and 50ug Streptomycin, per 500mL of media. Cells were initially diluted and distributed into wells of a 96 well plate at a dilution of 1/3 of a cell per well. Assuming a Poisson distribution, (formula in R: "dpois(# of expected cells per well,1/3)*96"), this results that on average ~23 wells would be expected to contain a single cell, and on average ~4 wells would be expected to contain two cells, and on average ~69 wells would be expected to contain no cells. After ~1-24 hours, once the cells attached to the surface of the plate, the wells were examined under a phase and under a fluorescent microscope to record wells that were initiated with a single cell. Conditioned, filtered media from CP-D cell cultures was used to facilitate clonal expansion when the clones are at low density. After cell growth reached near confluency within the wells of a 96-well plate, cells from the wells are trypsinized and transferred to a 6-well plate. Clones were passaged to larger plates as they grow so that they are never allowed to become confluent.

Cells are initially diluted and distributed into wells of a 96 well plate such that on average less than 35 wells contain cells and on average 10 out of the 35 wells contain a single cell. After 24 hours, once the cells have attached, the wells are examined under a light microscope using the objective lens set at 4X and 10X magnification to count the number of cells in each well of the 96 well plate. Conditioned, filtered media from CP-D cell cultures is used to facilitate clonal expansion when the clones are at low density. While the cells grow in a 96 well plate, they receive media change once per week. After 23 days, cells from the wells that were initiated with a single cell are trypsinized and transferred to a 6-well plate. At the 6-well plate stage, fresh KSFM media is used only and media is changed every 3-5 days. Clones are passaged to larger plates as they grow so that they are never allowed to become confluent.

I designed and followed the following protocol for preparing a DCA+KSFM solution:

1. A solution was made that contains both KSFM media and 250µM DCA. The final volume of this solution is 40mL. The molar mass of DCA is 392.6 g/mol, therefore to make 250µM DCA solution in 40mL, 3.926mg of DCA was measured (grams DCA= 0.04 L * 0.00025 mol/L * 392.6 g/mol). To measure such a tiny amount, I tared the scale instrument with an empty eppendorf tube, and then dispensed miniscule amounts of DCA powder into the tube to bring it to approximately 3.9mg. The final weight typically ranged between 3 and 5 mg due to the precision of dispensing such tiny amount of powder by hand into an eppendorf tube. At the 3-5mg extremes, the final solution would range between 191µM and 318µM, which was not far off the desired 250µM concentration.

2.The DCA powder was poured from the eppendorf tube into a beaker containing 40mL of KSFM. A magnetic stir bar and a hot plate (with the heat off) were used to dissolve the DCA into the solution, which typically took 40-50 mins of stirring. After dissolving, a filtered syringe was used to transfer the contents of the beaker to a 50mL plastic centrifuge tube. From that final 40mL 250µM DCA+KSFM solution, I prepared 0.1µM, 1µM, and 10µM solutions by dilution with fresh KSFM media, whenever cells needed fresh media, or whenever cells needed to be passaged from 96-well plates into 6-well plates and into larger plates.

5.2.1. Results and Discussion

When I tested concentrations of 50μM, 100μM, 250μM, and 500μM DCA on the single cell 96-well plates, all cells died within a few days, or, no cell growth occurred in that any single well can reach confluency.

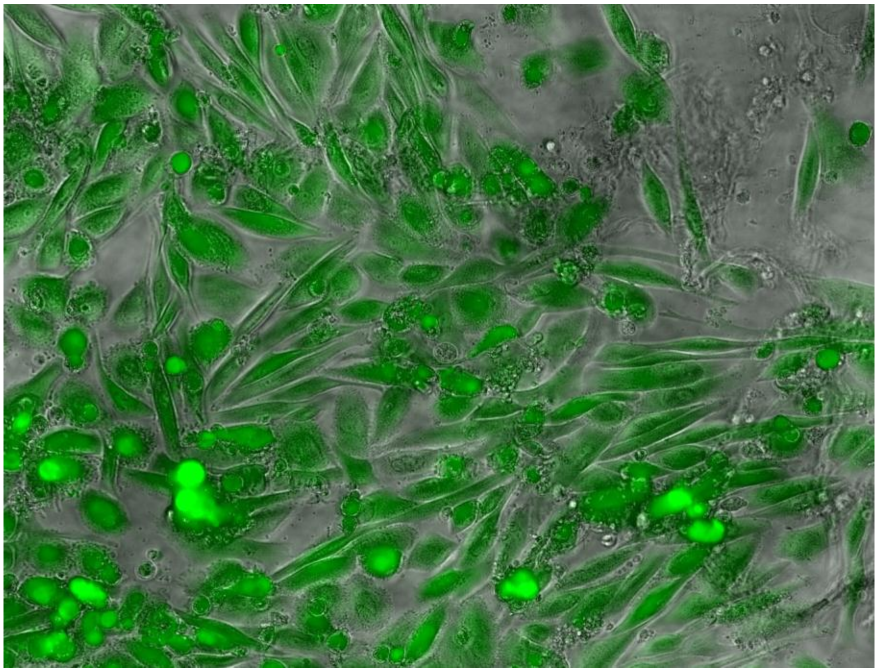


Figure 5.2. An example composite image from phase contrast and GFP fluorescence images under a 10x magnification of CP-D cells in a 6-well plate after several weeks of growth from a single starting cell.

CP-D cells have elongated sickle-like or ball-like shape and express high levels of green fluorescent protein (GFP). These images are taken several weeks after single cell cloning, i.e. all of the cells descent from a recent common ancestor cell that was seeded several weeks prior to the time of imaging.

DCA had a dose-related effect on inducing SGA in CP-D cells. The $\log_2$-transformed signal intensity R ratio between cloning round 2 versus cloning round 1 revealed differences between the rounds indicative of acquired SGAs. For example, taking the log2 ratio between round 1 and round 2 of the same clone under 1μM DCA revealed a SGA that was ~20Mb in size on chromosome 22 (Figure 5.3).

DCA appeared to induce microdeletions throughout the genome however rarely it seemed to be able to induce a large-scale SGA (Figure 5.3) similar to large scale SGAs seen in individuals a-m (Supplementary Figure 4.12) and in chromosome 9p of the single individual (Figure 2.6).

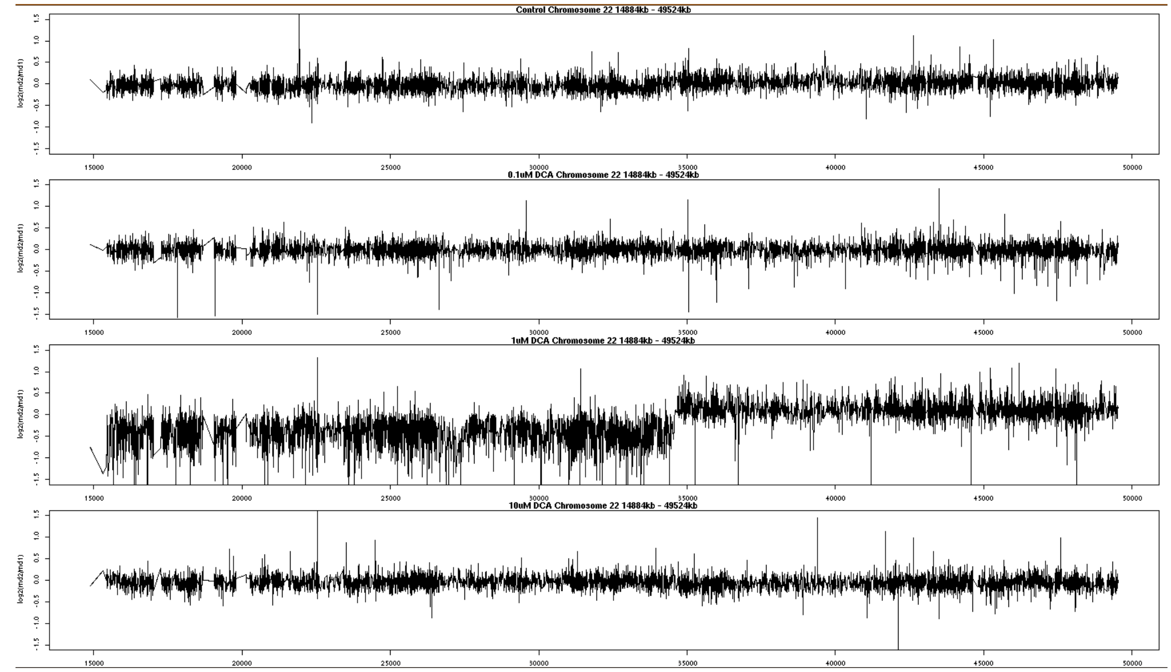


Figure 5.3. Comparing two rounds of single cell cloning revealed an acquired SGA that was approximately 20Mb in size on chromosome 22 in the clone that was grown under 1μM DCA (Panel C). In comparison, the clones grown under 0μM (control, panel A), 0.1μM (panel B), and 10μM DCA (panel D) show no apparent SGA at the same location. This result suggests that the bile acid DCA is capable of inducing large-scale SGA, as much as 20Mb in size, in a BE cell line *in vitro*. Importantly, the physiological doses of DCA in humans *in vivo* can be as much as 50-100μM [126].

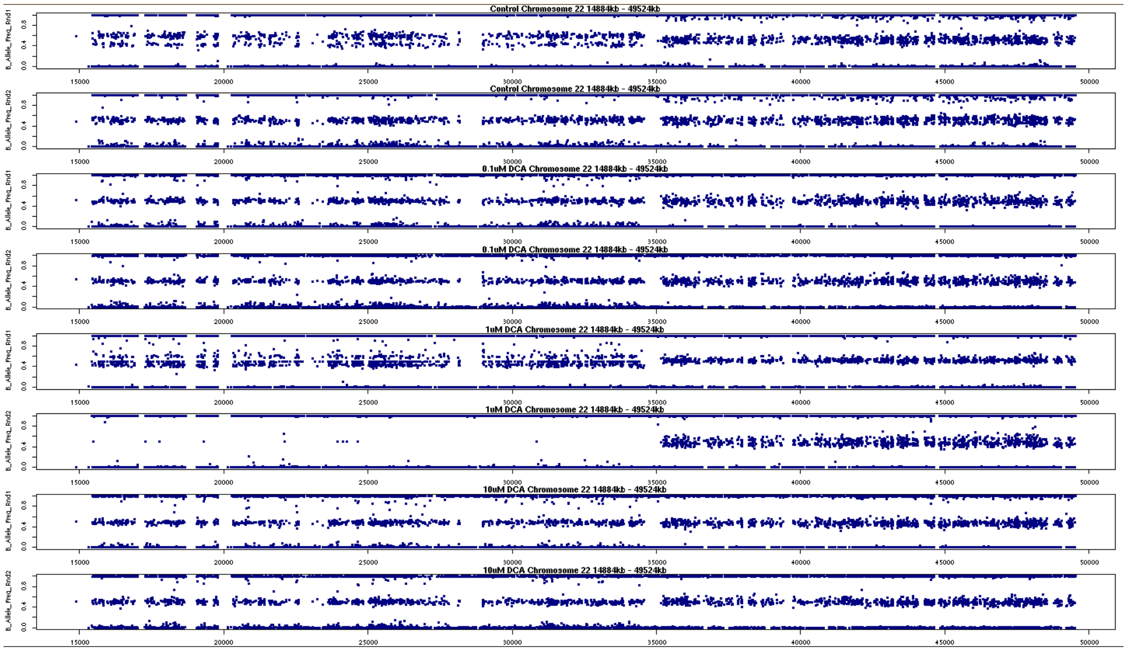


Figure 5.4. B allele frequency plots of all 8 samples from round 1 and round 2 of single cell cloning under 0μM (control, panels A,B), 0.1μM, 1μM and 10μM DCA exposure. Note that the 1μM DCA clone in round 2 acquired the ~20Mb single copy loss (Panel F) since no heterozygous SNPs are observed and the total signal intensity R is lower than the same clone froum round 1. (Figure 5.3, Panel C).

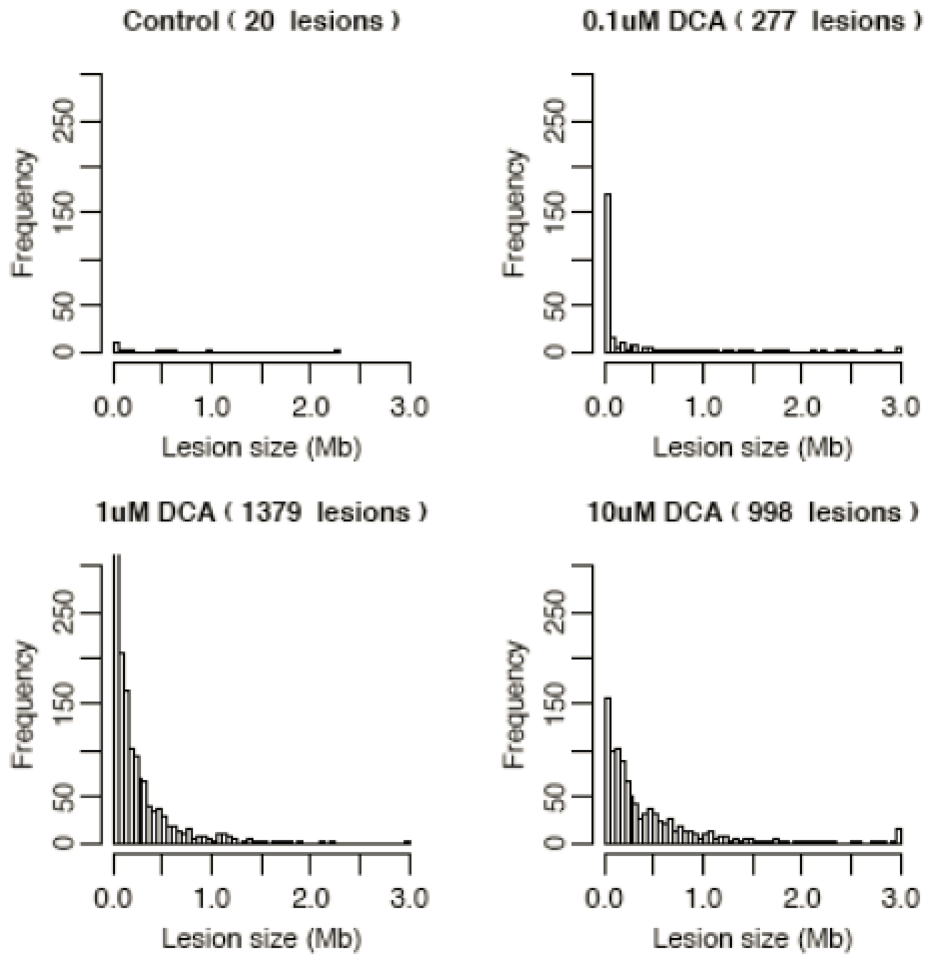


Figure 5.5. Histograms of the number and size of copy number alterations detected in evolving CP-D clones *in vitro* due to 0µM (control), 0.1µM, 1µM and 10µM DCA exposure to DCA, a genotoxic component of duodenal gastroesophageal reflux in individuals with Barrett's esophagus. Higher concentrations of DCA induce more and/or larger lesions.

I tested CP-D cells to determine if we could generate genetic lesions in culture. I used two rounds of single cell cloning to ensure that any differences between samples after the first and second round of cloning would be due to lesions that occurred during the course of the experiment, rather than due to sampling of different genotypes from a heterogeneous parental cell culture. I analyzed the results of Illumina 550K SNP arrays with the GLAD algorithm to identify regions of copy number differences between the first and second rounds of cloning. I tested exposures to a range of concentrations (0, 0.1, 1 and 10uM) of deoxycholic acid (DCA), a genotoxic component of gastroduodenal reflux in patients with BE. Figure 5.5 shows that the number and size of lesions we detected. Without a normal sample for comparison, LOH cannot be distinguished from constitutive homozygosity, so we have focused on copy number changes here.

Sample ID	Condition	96-well ID	Round
1	1uM DCA	B9	2
2	1uM DCA	F10	1
3	10uM DCA	C2	1
4	10uM DCA	B5	2
5	control	D12	1
6	control	F3	2
7	0.1 uM DCA	B7	1
8	0.1 uM DCA	D7	2

Table 5.2. Illumina 550k-SNP arrays were used to evaluate 8 samples for somatic abnormalities.

5.2.2. Conclusion

I showed that increasing dose of DCA incurred somatic genomic abnormalities in DNA of CP-D cells. This work contributed to successfully getting the American Cancer Society grant for measuring somatic mutation rate associated with NSAID exposure. NSAIDs exposure and DCA exposure may manifest in the same way in modulating the acquisition of somatic genomic abnormalities over time. Studying the modulating effect of NSAID exposure *in vitro* is difficult because it warrants constructing an immune response and inflammation in cell culture. Future studies in experimental cell culture evolution may further the development of computational methods of estimating the types and rates of somatic genomic abnormalities acquisition.

Chapter 6. Conclusions

6.1. Thesis

Epithelial esophageal adenocarcinoma is a disease of the somatic genome, the etiology of which is dependent on evolution by natural selection of somatic cells. I hypothesize that neoplastic cell populations in Barrett's esophagus maintain evolutionary stasis over decades despite having a genome, 1-3% of which is riddled with acquired somatic genomic abnormalities; and occasionally when massive amount (20% or more) of genomic abnormalities are detected in association with a local clonal expansion, the neoplastic cell populations can evolve and manifest malignancy. I also hypothesize that the puzzling phenomenon that 90-95% of individuals with Barrett's esophagus retain a benign course over time and never progress to esophageal adenocarcinoma is explained by the observations that evolutionary genomic stasis is maintained over decades in Barrett's neoplastic cell populations, which I have shown in one individual in Chapter 2 and in 12 individuals in Chapter 4. It has not escaped my attention that the same may hold true for other epithelial pre-malignant neoplastic conditions and their associated malignancies.

6.2. Future directions for research in evolution in cancer

My doctoral research is an example of applying methods in evolutionary biology to cancer data. While the title of this dissertation reads "evolutionary dynamics" I conclude with exposition of relative evolutionary stasis in Barrett's neoplastic cell populations. How does neoplastic evolution play out in Barrett's esophagus individuals over decades? While the predominant conceptual framework for neoplastic evolution is stepwise accumulation of genetic abnormalities [127], or gradualism model, our longitudinal observations in Barrett's individuals suggest stochastic bursts of acquisition of abnormalities and maintenance of the status quo in the majority of the genome, more consistent with a “punctuated equilibrium” model put forward for species-level evolution [128] . While our *in silico* modeling of genetic diversity predicts a monotonic increase over two decades, at the resolution of profiling the genomic state of biopsies and at the resolution of our sampling, our *in vivo* observations support maintenance of SGA amount which also translates to maintenance of genetic diversity with little relative change over time.

Massive somatic genomic abnormalities can be acquired in neoplastic cell populations, however, the *in vivo* evidence presented in Chapter 4 cannot distinguish conclusively whether a single cell mitosis or multiple cell mitoses induce massive amount of genomic abnormalities. There is an ongoing trial and error process during carcinogenesis, generating some cell lineages that persistently survive in neoplasms, and other cell lineages that only transiently survive and eventually die out outcompeted in the evolutionary race for survival. While conceptually, stages of carcinogenesis can be divided into initiation, promotion, and progression, the evolutionary process in neoplasm need not necessarily proceed in discreet steps, but rather manifest a continuum that is punctured by stochastic acquisition of SGAs. The *in vivo* evidence presented in Chapter 4, in my view, suggests a rugged fitness landscape where a massive SGA can elevate a cell to a stable fitness plateau sufficient to induce increase in numbers so that the cell transforms into a clonal population of cells. Any steps away from such plateaus or adaptive peaks tend to generate cells lineages having evolutionary dead ends, or not enough time is available for successful crossovers to other adaptive peaks in the genotype landscape. This is one plausible explanation of the apparent evolutionary stasis. Even in the massive SGA clone in individual j (Chapter 4, Figure 4.4) that clone had a mean SGA genotype of 588±18 Mb altered of the genome for a time interval of 3 years and being detected in 4 biopsies. That could be interpreted that the 588 Mb of SGAs that are approximately in common across the 4 biopsies form the genotypic configuration representing the adaptive peak of that clone within the fitness landscape. And the 18 Mb of standard deviation of SGA amount represents a tolerable deviation of ongoing trial and error in genotype space around that stable adaptive peak. Similarly, for most individuals, except b,j, and f, the detection of 53±30 Mb SGA amount could be interpreted as a stable adaptive peak of 53 Mb SGA, comprising mostly losses on chromosome 9 and fragile sites, and a 30 Mb of SGA trial and error around that adaptive peak, representing stochastic acquisition, retention, and purging of persistently or transiently surviving SGAs.

Is there an optimality criterion for a changed genome for manifesting persistent survival and clonal expansion? Certainly, baseline SGAs that are common across all 13 individuals (SGAs shown in black, Chapter 4, Supplementary Figure 4.12) are associated with the initial clonal expansion, whether or not they induced this initial clonal expansion that establishes the extent of the Barrett's segment. What is the minimal set of acquired SGAs that are required for an initial or a subsequent clonal expansion? Individual k had the minimum total amount of SGA with a maximum of 31 Mb SGA per biopsy. This is a relatively low amount of SGAs (0.97 % of 3,164 Mb-long human genome) associated with a Barrett's metaplasia. Could a set of genomically altered genes present within this 0.97% account for initiation of Barrett's metaplasia or is the initial clonal expansion driven by epigenetic modifications of the genome and/or the changed gastroesophageal reflux microenvironment? These questions warrant further larger-scale genomic studies that can resolve the etiology of Barrett's metaplasia and potentially advance understanding of disease genesis in other pre-malignant conditions in the gastrointestinal tract.

Do NSAIDs play a significant role in reducing SGA acquisition rate? While the effect of NSAIDs is difficult to investigate conditioned on the presence of evolutionary stasis, I hypothesize that NSAIDs interfere with promotion of clonal expansion of SGA-containing neoplastic cells. A mechanism of NSAID action may remain elusive of identification, because the NSAID effect manifests only over 5 or more years of use. It may be the long-term decrease of inflammation that potentially decreases secretion of growth factors of inflammatory cells present within Barrett's tissues that stimulate, or promote, clonal expansion of SGA-containing neoplastic stem cells triggering clonal growth driven by crypt fission.

All of the studies throughout the studies were based on DNA extracted from biopsies, in most cases processed by separating BE epithelial cells from stromal cells by an epithelial isolation technique. Assuming 6.5 picograms of DNA per diploid cell, 200ng of DNA corresponds to evaluating a mixture of the DNA of approximately 30,769 cells and for samples having lots of SGAs 200ng of DNA would alter the 30,769 estimate depending on the SGA loss to gain ratio. Therefore, since single cell genomics has not been feasible, all of my results and analyses are at the level of a biopsy, or a mean genotype of 30,000 cells. Further advancements in single cell genomics would allow the in-depth characterization of genetic diversity within a biopsy.

Future work could utilize approximate Bayesian computation approaches to match model-generated genetic diversity to the experimentally-observed genetic diversity to estimate the parameter values able to produce the observed dynamics. I hypothesize that a very low mutation rate of loci conferring higher crypt reproduction and that a very high reproduction rate-increasing effect of mutating such loci can produce a punctuated pattern of relative evolutionary stasis and rare bursts of massive SGAs that are also associated with local clonal expansions. I hypothesize that selective effects are probably very high, but it is very rare to achieve an aneuploid, massive SGA, clone that is stable and able to locally expand within the tissue.

In summary, the results in Chapter 4 will be of general interest to the scientific and medical community for the following reasons. Never before has neoplastic progression been studied in such detail, with 12+ biopsies over 5-8 time points and up to 19 years of follow-up within the same patients, using a whole genome assay. I found that aspirin and other NSAIDs, which are commonly available and cost effective medications, may exert their cancer preventive effect through lowering SGA rate. I found that the Barrett's segment in Barrett's esophagus individuals can remain in relative evolutionary stasis over decades of follow-up where only 0.6 and 7.8 SGAs may occur per biopsy, per genome, per year while on-NSAIDs and off-NSAIDs, respectively, suggesting an explanation for the low rate of progression for most Barrett's esophagus individuals. I demonstrated a new method for estimating somatic mutation (SGA) rates *in vivo,* in humans, which can be applied to any neoplasm with longitudinal samples (showing how evolutionary biologists can make fundamental contributions to cancer biology). I demonstrated a new, and fundamentally different, type of biomarker that measures the evolutionary dynamics of progression (SGA rate) not just the presence or absence of an abnormality. With longitudinal data spanning 6.4-19 years, I showed the genome-wide distribution patterns of SGAs over the evolutionary (natural) course of cancer development.

Finally, rapidly advancing sequencing technologies allow the measurement of the state of the entire genome. In the clinic, I would recommend that the detection of stable versus unstable genomes can help manage treatment options in individuals with Barrett's esophagus and other pre-malignant conditions. I would also recommend that NSAIDs may be used for reducing the rate of SGA acquisition in individuals with Barrett's esophagus. I believe it is feasible that future randomized controlled trials for cancer chemoprevention or prevention based on changes in diet and lifestyle could use measurements of evolutionary dynamics, for instance, changes in the level of somatic genomic abnormalities and phylogenetic tree shape imbalance both indicating emergence of malignant cell lineages, as intermediate endpoints of effectiveness.

Appendix

List of Figures

List of Tables

References

1. American Cancer Society (2012) Cancer Facts & Figures 2012. Atlanta.
2. Nowell PC (1976) The clonal evolution of tumor cell populations. Science (New York, NY) 194: 23–28.
3. Berenblum I, Shubik P (1947) A new, quantitative, approach to the study of the stages of chemical carcinogenesis in the mouse's skin. British journal of cancer 1: 383–391.
4. Foulds L (1964) Cellular Control Mechanisms and Cancer. In: Muhlbock O, Emmelot P, editors. Cellular Control Mechanisms and Cancer. Amsterdam: Elsevier. pp. 242–295.
5. Reid BJ, Li X, Galipeau PC, Vaughan TL (2010) Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. Nature reviews Cancer 10: 87–101. doi:10.1038/nrc2773.
6. Thomas T, Abrams KR, De Caestecker JS, Robinson RJ (2007) Meta analysis: Cancer risk in Barrett's oesophagus. Alimentary Pharmacology & Therapeutics 26: 1465–1477. doi:10.1111/j.1365-2036.2007.03528.x.
7. Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P (2011) Incidence of Adenocarcinoma among Patients with Barrett's Esophagus. New England Journal of Medicine 365: 1375–1383. doi:10.1056/NEJMoa1103042.
8. Reid BJ, Kostadinov R, Maley CC (2011) New Strategies in Barrett's Esophagus: Integrating clonal evolutionary theory with clinical management. Clinical cancer research : an official journal of the American Association for Cancer Research 17: 3512–3519. doi:10.1158/1078-0432.CCR-09-2358.
9. von Zglinicki T (2002) Oxidative stress shortens telomeres. Trends in biochemical sciences 27: 339–344.
10. Jenkins GJS, Cronin J, Alhamdani A, Rawat N, D'Souza F, et al. (2008) The bile acid deoxycholic acid has a non-linear dose response for DNA damage and possibly NF-kappaB activation in oesophageal cells, with a mechanism of action involving ROS. Mutagenesis 23: 399–405. doi:10.1093/mutage/gen029.
11. Grisham MB, Jourd'heuil D, Wink DA (2000) Review article: chronic inflammation and reactive oxygen and nitrogen metabolism--implications in DNA damage and mutagenesis. Alimentary pharmacology & therapeutics 14 Suppl 1: 3–9.
12. Sihvo EIT, Salminen JT, Rantanen TK, Rämö OJ, Ahotupa M, et al. (2002) Oxidative stress has a role in malignant transformation in Barrett's oesophagus. International journal of cancer Journal international du cancer 102: 551–555. doi:10.1002/ijc.10755.
13. Trayhurn P, Bing C, Wood IS (2006) Adipose tissue and adipokines--energy regulation from the human perspective. The Journal of nutrition 136: 1935S–1939S.
14. Turker MS, Gage BM, Rose JA, Elroy D, Ponomareva ON, et al. (1999) A novel signature mutation for oxidative damage resembles a mutational pattern found commonly in human cancers. Cancer research 59: 1837–1839.
15. Orlando RC (2007) Mucosal Defense in Barrett's Esophagus, in Barrett's Esophagus and Esophageal Adenocarcinoma, Second Edition (eds P. Sharma and R. Sampliner). Sharma P, Sampliner R, editors Oxford, UK: Blackwell Publishing Ltd. p. doi:10.1002/9780470987513.

16. Tobey NA, Argote CM, Vanegas XC, Barlow W, Orlando RC (2007) Electrical parameters and ion species for active transport in human esophageal stratified squamous epithelium and Barrett's specialized columnar epithelium. American journal of physiology Gastrointestinal and liver physiology 293: G264–70. doi:10.1152/ajpgi.00047.2007.

17. Cairns J (1975) Mutation selection and the natural history of cancer. Nature 255: 197–200. doi:10.1038/255197a0.

18. Levine DS, Rubin CE, Reid BJ, Haggitt RC (1989) Specialized metaplastic columnar epithelium in Barrett's esophagus. A comparative transmission electron microscopic study. Laboratory investigation; a journal of technical methods and pathology 60: 418–432.

19. Levine DS, Reid BJ, Haggitt RC, Rubin CE, Rabinovitch PS (1989) Correlation of ultrastructural aberrations with dysplasia and flow cytometric abnormalities in Barrett's epithelium. Gastroenterology 96: 355–367.

20. Dixon J, Strugala V, Griffin SM, Welfare MR, Dettmar PW, et al. (2001) Esophageal mucin: an adherent mucus gel barrier is absent in the normal esophagus but present in columnar-lined Barrett's esophagus. The American journal of gastroenterology 96: 2575–2583. doi:10.1111/j.1572-0241.2001.04159.x.

21. Glickman JN, Blount PL, Sanchez CA, Cowan DS, Wongsurawat VJ, et al. (2006) Mucin core polypeptide expression in the progression of neoplasia in Barrett's esophagus. Human pathology 37: 1304–1315. doi:10.1016/j.humpath.2006.03.023.

22. Lao-Sirieix P, Corovic A, Jankowski J, Lowe A, Triadafilopoulos G, et al. (2008) Physiological and molecular analysis of acid loading mechanisms in squamous and columnar-lined esophagus. Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / ISDE 21: 529–538. doi:10.1111/j.1442-2050.2007.00807.x.

23. Jovov B, Van Itallie CM, Shaheen NJ, Carson JL, Gambling TM, et al. (2007) Claudin-18: a dominant tight junction protein in Barrett's esophagus and likely contributor to its acid resistance. American journal of physiology Gastrointestinal and liver physiology 293: G1106–13. doi:10.1152/ajpgi.00158.2007.

24. Ostrowski J, Mikula M, Karczmarski J, Rubel T, Wyrwicz LS, et al. (2007) Molecular defense mechanisms of Barrett's metaplasia estimated by an integrative genomics. Journal of molecular medicine (Berlin, Germany) 85: 733–743. doi:10.1007/s00109-007-0176-3.

25. Willis R (1952) The spread of tumors in the human body. London: Butterworth and Co., Ltd. p.

26. Dorland WAN (2011) Dorland's illustrated medical dictionary. 31st ed. W.B. Saunders Co. p.

27. Merlo LMFF, Pepper JW, Reid BJ, Maley CC (2006) Cancer as an evolutionary and ecological process. Nat Rev Cancer 6: 924–935. doi:10.1038/nrc2013.

28. Hanahan D, Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. Cell 144: 646–674. doi:10.1016/j.cell.2011.02.013.

29. Boveri T (1914) Zur Frage der Entstehung maligner Tumoren. Jena: Gustav Fischer. p.

30. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, et al. (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nature Genetics 38: 468–473. doi:10.1038/ng1768.
31. Merlo LMF, Maley CC (2010) The role of genetic diversity in cancer. The Journal of Clinical Investigation 120: 401–403. doi:10.1172/JCI42088.
32. Park SY, Lee HE, Li H, Shipitsin M, Gelman R, et al. (2010) Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research 16: 876–887. doi:10.1158/1078-0432.CCR-09-1532.
33. Park SY, Gönen M, Kim HJ, Michor F, Polyak K (2010) Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. The Journal of clinical investigation 120: 636–644. doi:10.1172/JCI40724.
34. Leedham SJ, Preston SL, McDonald SAC, Elia G, Bhandari P, et al. (2008) Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus. Gut 57: 1041–1048. doi:10.1136/gut.2007.143339.
35. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905. doi:10.1038/nature08822.
36. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. Nature 396: 643–649. doi:10.1038/25292.
37. Rajagopalan H, Lengauer C (2004) Aneuploidy and cancer. Nature 432: 338–341. doi:10.1038/nature03099.
38. Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C (2003) The significance of unstable chromosomes in colorectal cancer. Nature reviews Cancer 3: 695–701. doi:10.1038/nrc1165.
39. Nowak MA, Komarova NL, Sengupta A, Jallepalli PV, Shih I-M, et al. (2002) The role of chromosomal instability in tumor initiation. Proceedings of the National Academy of Sciences of the United States of America 99: 16226–16231. doi:10.1073/pnas.202617399.
40. Vogelstein B, Fearon ER, Hamilton SR, Feinberg AP (1985) Use of restriction fragment length polymorphisms to determine the clonal origin of human tumors. Science (New York, NY) 227: 642–645.
41. Navin NE, Hicks J (2010) Tracing the tumor lineage. Molecular oncology 4: 267–283. doi:10.1016/j.molonc.2010.04.010.
42. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. Genome research 20: 68–80. doi:10.1101/gr.099622.109.
43. Salipante SJ, Horwitz MS (2007) A phylogenetic approach to mapping cell fate. Current topics in developmental biology 79: 157–184. doi:10.1016/S0070-2153(06)79006-8.
44. Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E (2005) Genomic variability within an organism exposes its cell lineage tree. PLoS computational biology 1: e50. doi:10.1371/journal.pcbi.0010050.

45. Wasserstrom A, Frumkin D, Adar R, Itzkovitz S, Stern T, et al. (2008) Estimating cell depth from somatic mutations. PLoS computational biology 4: e1000058. doi:10.1371/journal.pcbi.1000058.

46. Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, et al. (2008) Cell lineage analysis of a mouse tumor. Cancer Research 68: 5924–5931. doi:10.1158/0008-5472.CAN-07-6216.

47. Carlson CA, Kas A, Kirkwood R, Hays LE, Preston BD, et al. (2012) Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nature methods 9: 78–80. doi:10.1038/nmeth.1781.

48. Louhelainen J, Wijkström H, Hemminki K (2000) Allelic losses demonstrate monoclonality of multifocal bladder tumors. International journal of cancer Journal international du cancer 87: 522–527.

49. Ruiz C, Lenkiewicz E, Evers L, Holley T, Robeson A, et al. (2011) Advancing a clinically relevant perspective of the clonal nature of cancer. Proceedings of the National Academy of Sciences of the United States of America 108: 12054–12059. doi:10.1073/pnas.1104009108.

50. Salk JJ, Horwitz MS (2010) Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Seminars in Cancer Biology. doi:10.1016/j.semcancer.2010.10.008.

51. Salk JJ, Salipante SJ, Risques RA, Crispin DA, Li L, et al. (2009) Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences of the United States of America 106: 20871–20876. doi:10.1073/pnas.0909428106.

52. Tsao JL, Yatabe Y, Salovaara R, Järvinen HJ, Mecklin JP, et al. (2000) Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences of the United States of America 97: 1236–1241.

53. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, et al. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proceedings of the National Academy of Sciences 105: 13081–13086. doi:10.1073/pnas.0801523105.

54. Yachida S, Jones S, Bozic I, Antal T, Leary R, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467: 1114–1117. doi:10.1038/nature09515.

55. Chao DL, Eck JT, Brash DE, Maley CC, Luebeck EG (2008) Preneoplastic lesion growth driven by the death of adjacent normal stem cells. Proceedings of the National Academy of Sciences 105: 15034 –15039. doi:10.1073/pnas.0802211105.

56. Martens E a, Kostadinov R, Maley CC, Hallatschek O (2011) Spatial structure increases the waiting time for cancer. New Journal of Physics 13: 115014. doi:10.1088/1367-2630/13/11/115014.

57. Graham TA, McDonald SAC (2010) Genetic diversity during the development of Barrett's oesophagus-associated adenocarcinoma: how, when and why? Biochemical Society Transactions 38: 374. doi:10.1042/BST0380374.

58. Sottoriva A, Vermeulen L, Tavaré S (2011) Modeling Evolutionary Dynamics of Epigenetic Mutations in Hierarchically Organized Tumors. PLoS Computational Biology 7: e1001132. doi:10.1371/journal.pcbi.1001132.

59. Nicolas P, Kim K-M, Shibata D, Tavaré S (2007) The Stem Cell Population of the Human Colon Crypt: Analysis via Methylation Patterns. PLoS Computational Biology 3. doi:10.1371/journal.pcbi.0030028.
60. F Kingman J (2000) Origins of the coalescent. 1974-1982. Genetics 156: 1461.
61. Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of Single Nucleotide Polymorphism Data for Estimating Population Parameters. Genetics 156: 439–447.
62. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161: 1307–1320.
63. Excoffier L, Novembre J, Schneider S (n.d.) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. The Journal of heredity 91: 506–509.
64. Rothwell PM, Fowkes FGR, Belch JFF, Ogawa H, Warlow CP, et al. (2011) Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. Lancet 377: 31–41. doi:10.1016/S0140-6736(10)62110-1.
65. Corley DA, Kerlikowske K, Verma R, Buffler P (2003) Protective association of aspirin/NSAIDs and esophageal cancer: a systematic review and meta-analysis. Gastroenterology 124: 47–56. doi:10.1053/gast.2003.50008.
66. Abnet CC, Freedman ND, Kamangar F, Leitzmann MF, Hollenbeck AR, et al. (2009) Non-steroidal anti-inflammatory drugs and risk of gastric and oesophageal adenocarcinomas: results from a cohort study and a meta-analysis. British journal of cancer 100: 551–557. doi:10.1038/sj.bjc.6604880.
67. Vaughan TL, Dong LM, Blount PL, Ayub K, Odze RD, et al. (2005) Non-steroidal anti-inflammatory drugs and risk of neoplastic progression in Barrett's oesophagus: a prospective study. The Lancet Oncology 6: 945–952. doi:10.1016/S1470-2045(05)70431-9.
68. Galipeau PC, Li X, Blount PL, Maley CC, Sanchez CA, et al. (2007) NSAIDs modulate CDKN2A, TP53, and DNA content risk for progression to esophageal adenocarcinoma. PLoS medicine 4: e67. doi:10.1371/journal.pmed.0040067.
69. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921. doi:10.1038/35057062.
70. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science (New York, NY) 291: 1304–1351. doi:10.1126/science.1058040.
71. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, et al. (2004) Decoding randomly ordered DNA arrays. Genome research 14: 870–877. doi:10.1101/gr.2255804.
72. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Research 16: 1136–1148. doi:10.1101/gr.5402306.
73. Cheng H, Bjerknes M, Amar J (1984) Methods for the determination of epithelial cell kinetic parameters of human colonic epithelium isolated from surgical and biopsy specimens. Gastroenterology 86: 78–85.
74. Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c.

75. Diskin SJ, Li M, Hou C, Yang S, Glessner J, et al. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Research 36: e126. doi:10.1093/nar/gkn556.

76. Lai LA, Kostadinov R, Barrett MT, Peiffer DA, Pokholok D, et al. (2010) Deletion at Fragile Sites Is a Common and Early Event in Barrett's Esophagus. Molecular cancer research 8: 1084–1094. doi:10.1158/1541-7786.MCR-09-0529.

77. Paulson TG, Reid BJ (2004) Focus on Barrett's esophagus and esophageal adenocarcinoma. Cancer cell 6: 11–16. doi:10.1016/j.ccr.2004.06.021.

78. Li X, Galipeau PC, Sanchez CA, Blount PL, Maley CC, et al. (2008) Single nucleotide polymorphism-based genome-wide chromosome copy change, loss of heterozygosity, and aneuploidy in Barrett's esophagus neoplastic progression. Cancer prevention research (Philadelphia, Pa) 1: 413–423. doi:10.1158/1940-6207.CAPR-08-0121.

79. Paulson TG, Maley CC, Li X, Li H, Sanchez CA, et al. (2009) Chromosomal instability and copy number alterations in Barrett's esophagus and esophageal adenocarcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research 15: 3305–3314. doi:10.1158/1078-0432.CCR-08-2494.

80. Hupé P, Stransky N, Thiery J-P, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics (Oxford, England) 20: 3413–3422. doi:10.1093/bioinformatics/bth418.

81. Hsu L, Self SG, Grove D, Randolph T, Wang K, et al. (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics (Oxford, England) 6: 211–226. doi:10.1093/biostatistics/kxi004.

82. Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, et al. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome research 16: 1149–1158. doi:10.1101/gr.5076506.

83. R Development Core Team (2011) R: A Language and Environment for Statistical Computing.

84. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome research 19: 1639–1645. doi:10.1101/gr.092759.109.

85. Wan L, Fu WJ, Deng M, Qian M (2008) A Method to Correct Systematic Bias in Affymetrix SNP Arrays. International Conference on Biomedical Engineering and Informatics. IEEE Computer Society. pp. 442–446. doi:10.1109/BMEI.2008.41.

86. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, et al. (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC genomics 10: 588. doi:10.1186/1471-2164-10-588.

87. Heng HHQ, Stevens JB, Bremer SW, Liu G, Abdallah BY, et al. (2011) Evolutionary mechanisms and diversity in cancer. Advances in cancer research 112: 217–253. doi:10.1016/B978-0-12-387688-1.00008-9.

88. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, et al. (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nature genetics 38: 468–473. doi:10.1038/ng1768.

89. Merlo LMF, Shah NA, Li X, Blount PL, Vaughan TL, et al. (2010) A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of

progression to esophageal adenocarcinoma. Cancer prevention research (Philadelphia, Pa) 3: 1388–1397. doi:10.1158/1940-6207.CAPR-10-0108.

90. Galipeau PC, Prevo LJ, Sanchez CA, Longton GM, Reid BJ (1999) Clonal Expansion and Loss of Heterozygosity at Chromosomes 9p and 17p in Premalignant Esophageal (Barrett's) Tissue. JNCI Journal of the National Cancer Institute 91: 2087–2095. doi:10.1093/jnci/91.24.2087.

91. Herrero-Jimenez P, Tomita-Mitchell A, Furth EE, Morgenthaler S, Thilly WG (2000) Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. Mutation Research 447: 73–116.

92. Meza R, Hazelton WD, Colditz GA, Moolgavkar SH (2008) Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. Cancer causes & control : CCC 19: 317–328. doi:10.1007/s10552-007-9094-5.

93. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, et al. (2007) Genetic Progression and the Waiting Time to Cancer. PLoS Comput Biol 3: e225. doi:10.1371/journal.pcbi.0030225.

94. Totafurno J, Bjerknes M, Cheng H (1987) The crypt cycle. Crypt and villus production in the adult intestinal epithelium. Biophysical Journal 52: 279–294. doi:10.1016/S0006-3495(87)83215-0.

95. Greaves LC, Preston SL, Tadrous PJ, Taylor RW, Barron MJ, et al. (2006) Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. Proceedings of the National Academy of Sciences of the United States of America 103: 714–719. doi:10.1073/pnas.0505903103.

96. McDonald SAC, Preston SL, Greaves LC, Leedham SJ, Lovell MA, et al. (2006) Clonal expansion in the human gut: mitochondrial DNA mutations show us the way. Cell Cycle (Georgetown, Tex) 5: 808–811.

97. Cheng H, Matthew Bjerknes, Jack Amar, Geoffrey Gardiner (1986) Crypt production in normal and diseased human colonic epithelium. The Anatomical Record 216: 44–48. doi:10.1002/ar.1092160108.

98. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, et al. (2011) Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 144: 27–40. doi:10.1016/j.cell.2010.11.055.

99. Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, et al. (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome research 16: 1465–1479. doi:10.1101/gr.5460106.

100. Heyer W-D, Ehmsen KT, Liu J (2010) Regulation of homologous recombination in eukaryotes. Annual review of genetics 44: 113–139. doi:10.1146/annurev-genet-051710-150955.

101. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481: 306–313. doi:10.1038/nature10762.

102. Li X, Blount PL, Vaughan TL, Reid BJ (2011) Application of biomarkers in cancer risk management: evaluation from stochastic clonal evolutionary and dynamic system

optimization points of view. PLoS computational biology 7: e1001087. doi:10.1371/journal.pcbi.1001087.

103. Spechler SJ, Fitzgerald RC, Prasad GA, Wang KK (2010) History, molecular mechanisms, and endoscopic treatment of Barrett's esophagus. Gastroenterology 138: 854–869. doi:10.1053/j.gastro.2010.01.002.

104. Hartwell L (1992) Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. Cell 71: 543–546. doi:10.1016/0092-8674(92)90586-2.

105. Rabinovitch PS, Reid BJ, Haggitt RC, Norwood TH, Rubin CE (1988) Progression to cancer in Barrett's esophagus is associated with genomic instability. Laboratory investigation; a journal of technical methods and pathology 60: 65–71.

106. Liao LM, Vaughan TL, Corley DA, Cook MB, Casson AG, et al. (2011) Non-Steroidal Anti-Inflammatory Drug Use Reduces Risk for Adenocarcinomas of the Esophagus and Esophagogastric Junction in a Pooled Analysis. Gastroenterology. doi:10.1053/j.gastro.2011.11.019.

107. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7: 214. doi:10.1186/1471-2148-7-214.

108. Orlando RC (2010) The integrity of the esophageal mucosa. Balance between offensive and defensive mechanisms. Best practice & research Clinical gastroenterology 24: 873–882. doi:10.1016/j.bpg.2010.08.008.

109. Orlando RC (2006) Esophageal mucosal defense mechanisms. GI Motility online. doi:10.1038/gimo15.

110. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature: 1–6. doi:10.1038/nature09807.

111. Levine DS, Haggitt RC, Blount PL, Rabinovitch PS, Rusch VW, et al. (1993) An endoscopic biopsy protocol can differentiate high-grade dysplasia from early adenocarcinoma in Barrett's esophagus. Gastroenterology 105: 40–50.

112. Farrow D, Vaughan T, Hansten P, Stanford J, Risch H, et al. (1998) Use of aspirin and other nonsteroidal anti-inflammatory drugs and risk of esophageal and gastric cancer. Cancer Epidemiol Biomarkers Prev 7: 97–102.

113. Ben-Yaacov E, Eldar YC (2008) A fast and flexible method for the segmentation of aCGH data. Bioinformatics (Oxford, England) 24: i139–45. doi:10.1093/bioinformatics/btn272.

114. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 39: 906–913. doi:10.1038/ng2088.

115. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534.

116. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution 17: 368–376.

117. Swofford DL (2002) PAUP*.Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts. doi:10.1007/BF02198856.

118. Wang X, Ouyang H, Yamamoto Y, Kumar PA, Wei TS, et al. (2011) Residual Embryonic Cells as Precursors of a Barrett's-like Metaplasia. Cell 145: 1023–1035. doi:10.1016/j.cell.2011.05.026.

119. Johns B a E (1952) Developmental changes in the oesophageal epithelium in man. Journal of anatomy 86: 431–442.

120. Drewes A, Pedersen J, Liu W, Arendt-Nielsen L, Gregersen H (2003) Controlled mechanical distension of the human oesophagus: sensory and biomechanical findings. Scandinavian journal of gastroenterology 38: 27–35.

121. Mooers A, Heard S (1997) Inferring Evolutionary Process from Phylogenetic Tree Shape. The Quarterly Review of Biology 72: 31 - 54.

122. Bortolussi N, Durand E, Blum M, François O (2006) apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics (Oxford, England) 22: 363–364. doi:10.1093/bioinformatics/bti798.

123. Steel GG (1968) Cell loss from experimental tumors. Cell Proliferation 1: 193–207. doi:10.1111/j.1365-2184.1968.tb00318.x.

124. Cummins AG, Catto-Smith AG, Cameron DJ, Couper RT, Davidson GP, et al. (2008) Crypt fission peaks early during infancy and crypt hyperplasia broadly peaks during infancy and childhood in the small intestine of humans. Journal of pediatric gastroenterology and nutrition 47: 153–157. doi:10.1097/MPG.0b013e3181604d27.

125. Felsenstein J (2003) Inferring Phylogenies. Sinauer Associates. p.

126. Jenkins GJS, D'Souza FR, Suzen SH, Eltahir ZS, James SA, et al. (2007) Deoxycholic acid at neutral and acid pH, is genotoxic to oesophageal cells through the induction of ROS: The potential role of anti-oxidants in Barrett's oesophagus. Carcinogenesis 28: 136–142. doi:10.1093/carcin/bgl147.

127. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. Cell 61: 759–767.

128. Gould SJ, Eldredge N (1993) Punctuated equilibrium comes of age. Nature 366: 223–227. doi:10.1038/366223a0.