Development and Application of Novel Methods to Study Tumor Heterogeneity and Cancer Genome Evolution

by

Bo Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2014

Doctoral Committee:

Associate Professor Jun Li, Chair
Professor Daniel M. Burns Jr.
Professor Eric R. Fearon
Professor Kerby A. Shedden
Associate Professor Sebastian K. Zoellner

To my wife and my parents

ACKNOWLEDGEMENTS

I am deeply indebted to my thesis advisor Dr. Jun Li. Throughout my graduate school, Dr. Jun Li not only guided me through difficult scientific researches with patience, but also affected me by his careful and responsible personality. As a mentor, Dr. Jun Li is very strict on my researches. He would impose harsh critics for me to address and challenge my discoveries. At first, I was very struggling over his tough comments, but gradually, I started to appreciate his rigorous attitude towards scientific discoveries and truly objective point of views. I learned to think in the same way and be extremely careful with my findings. He also gave me enough freedom to explore my true interest and trained me to work and think independently. So far, I am so grateful to have him as my advisor and I cannot say how much I have benefit from his mentorship.

I also owe my thankfulness to the other members of my dissertation committee, Drs. Daniel Burns, Eric Fearon, Kerby Shedden and Sebastian Zoellner, for their guidance during my research. Dr. Burns has been both in my preliminary exam committee and thesis committee. He is very insightful and encouraged me to tackle difficult challenges in mathematical modeling and systematical thinking. As the only cancer biologist, Dr. Fearon gave me lots of useful insights that I could not obtain from my other members of my committee. I have known Kerby since the year 2009. He is my good friend and teacher. The statistical expertise

he shared with me helped me to address reviewer's comments. Dr. Zoellner is a great lecturer. I benefited from his teaching and then fortunately to have him join my thesis committee. His expertise in statistical genetics and next generation sequencing greatly helped my short tandem repeat genotyper project.

Besides my committee, I would like to thank Dr. Margit Burmeister, who has recruited me to the program of bioinformatics in University of Michigan, and has been a great mentor and friend ever since. She helped me to select my courses wisely, taught me about the culture differences while I was first in the US, comforted me as a friend when I was in trouble and also collaborated with me in an exciting project of short tandem repeat expansion. I am also grateful to Dr. Yuan Zhu and his postdoctoral fellow, Dr. Yinghua Li for their productive and insightful discussions in another collaboration studying brain cancer using genetic engineered mouse models.

Next I want to thank my previous and present lab members for the joyful days I spent and the tough times I have been through together. In particular, I owe my thanks to Weiping Peng, who has been such a good friend for both my wife and me. Also, I would like to thank Yasin Senbabaoglu for him teaching me how to drive, Jishu Xu both for her professional help in my work and useful suggestions in my life, Bilge Ozel who cooked the best mini blue-berry muffins and Yu-yu Ren for his practical suggestions on areas that I am not familiar with.

I owe special thanks to my beloved wife, Lin Zhong, for the good times and tough times we have been through together. It is impossible for me to finish this journey without her accompany, loving cares and encouragements. I also want to thank my parents and parent-in-laws. Although we are separate most of the time, I know they are always my help

and support.

Last, I would like to thank my collaborators, Yasin Senbabaoglu, Weiping Peng, Jishu Xu and Min-Lee Yang who have contributed to my publication of **Chapter 2**, for data pre-processing, analysis, and useful discussions during manuscript preparation.

Table of Contents

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

ABSTRACT

Cancer is one of the leading causes of death worldwide. In recent years, with the aid of high-throughput genomic technologies, large cohorts of tumor samples have been analyzed to characterize molecular aberrations in many cancer types. These studies have generated enormous amount of cancer genomics data, providing not only new opportunities to understand tumor evolution and cancer progression mechanisms but also new challenges in efficiently and rigorously analyzing the data. Heterogeneity is an important feature of cancer and has significant impact on the diagnosis and treatment of the disease. My dissertation focuses on developing new bioinformatics and biostatistical approaches to study the heterogeneity and evolutionary history of cancer genomes. Under this theme, my thesis consists of four main chapters. First, I have developed an algorithm to infer aneuploid and euploid cell mixing ratios using allele-specific DNA copy number alteration (CNA) data, and made a striking discovery that gene expression patterns in brain and ovarian tumors are strongly influenced by aneuploid content. The ability to infer mixing ratios allowed me to revise the current classification system for glioblastoma, with better predictive power of clinical outcome than previous results. Second, I developed a Clonal Heterogeneity Analysis Tool (CHAT) that estimates cellular fractions for individual CNAs and individual somatic mutations, allowing us to use the distribution of these fractions to inform the macroscopic clonal architecture and the relative order of occurrence of somatic changes. For example, a

CNA with a higher frequency in the cell population may have occurred earlier in tumor development or conferred a greater growth rate, therefore is more likely to contain driver genes. Third, I developed a method to detect short tandem repeat (STR) variation using paired-end short-read next-generation DNA sequencing data. Unlike previous methods which are limited to finding short STR alleles, my method is capable of finding both STR alleles shorter than a read and those longer than the read or the read pair (i.e., the insert size of the library). This capability addresses the need to reliably detect expanded STR alleles in germline DNA that underlie many rare inherited diseases as well as somatic aberrations characterized by microsatellite instability. In sum, my dissertation work led to the development of several new methods to study tumor heterogeneity. Their applications in multiple tumor types have made important contributions in understanding the mechanisms of tumor evolution. The work of my thesis is not only helpful to study the nature of tumor heterogeneity and evolution, but also provides a basis for assessing the impact of diverse tumor cell population on clinically important aspects such as subtype classification, prognosis and therapeutic resistance.

Chapter 1. Introduction

1.1 Background

1.1.1 Significance of Tumor Heterogeneity

Cancer is a leading cause of death in the US and worldwide (Jemal et al., 2008). Between 100 and 200 billion dollars are spent each year in US alone on cancer patient care (National Cancer Institute: http://costprojections.cancer.gov/). Despite decades of intensive efforts, it remains difficult to provide early diagnosis or effective treatment to many types of cancers. Most advanced cancers remain incurable. Oncologists commonly apply chemotherapies or radiotherapies to kill proliferating cancer cells. Nonetheless, eradication of tumor cells via cytotoxic therapies is rarely complete, and in most scenarios tumor eventually develops drug resistance, i.e., becomes insensitive to the treatment (Marusyk and Polyak, 2010). Emerging evidence has linked the resilience of cancers to its intrinsic heterogeneity, suggesting that a tumor can survive multiple environmental challenges, including immune response, inflammatory stimuli, clinical treatment etc., due to its large pool of genetically divergent cells that allow rapid adaptation and continued evolution (Yates and Campbell, 2012, Marte, 2013). Understanding such adaptation and evolution is a major focus of today's cancer

research.

Tumor heterogeneity refers to genetic and functional differences between cells in a tumor cell population (Heppner, 1984). Historically, studies on this topic dated back to 1950s when histologists dissected ascites tumors in mice and observed uneven chromosome numbers among individual tumor cells under microscope (Levan and Hauschka, 1953). In the past decade, with the development of high-throughput technologies, the paradigm of cancer research has shifted from physiological manifestations into mechanistic study of the underpinning molecular signatures of tumor heterogeneity (Marusyk et al., 2012). These studies have provided a progressively deeper understanding of this complex human disease.

1.1.2 Levels of Tumor Heterogeneity

Heterogeneity in tumors can be organized according to different levels of biological organization: **inter-tumor heterogeneity**, which refers to the differences among tumors between patients or within the same patient; and **intra-tumor heterogeneity**, which includes both tumor/normal cell mixing and potential co-existence of multiple subclones in the tumor cell population.

Inter-tumor heterogeneity refers to the genetic and phenotypic differences between individual tumors. This heterogeneity includes inter-patient differences, including how cancer patients have distinct clinical outcomes, such as responding differently to the same therapies. Such inter-patient differences reflect genetic, developmental, and environmental differences. For

example, breast cancers can be categorized by grade, by recurrent mRNA expression patterns or characteristic mutations, or by the presence and absence of certain hormone receptors. The classification of a patient's cancer into discrete subtypes—by clinical and/or cellular and molecular features—is an important area of research and patient care. In the example of breast cancers, the triple-negative subtype is defined by the absence of the estrogen receptor, progesterone receptor, and the her2 receptor. This subtype is more aggressive than others and leads to the worst prognosis. Studies of inter-patient heterogeneity require a large sample size to adequately discover recurrent events and subtypes. Many of these studies do not consider intra-tumor heterogeneity, rather they treat each patients tumor sample as a uniform entity, containing a homogeneous population of cells.

Inter-tumor heterogeneity also includes differences among tumors within a cancer patient (Vogelstein et al., 2013), both among primary tumors and between primary and metastatic tumors. The metastatic tumors, while originated from the primary tumor, may have acquired additional molecular aberrations and may be different from each other as they adapt to distinct local environments (Yachida et al., 2010).

Intra-tumor heterogeneity could simply involve different levels of Tumor-Normal mixing. A surgically obtained tumor specimen could contain many cell types, including tumor cells, surrounding stromal cells, blood vessels, and infiltrating immune cells. Although many studies have applied extensive sample selection according to histological tumor "purity", these procedures cannot completely remove the admixture of normal non-cancerous cells; and as a result, the data may not derive solely from the tumor cell populations, but may include the contribution of other cells.

Even when the sample contains 100% tumor cells, these cells may belong to different tumor subclone, adding another layer of intra-tumor heterogeneity. Even 1 mm^3 of tissue material may contain millions of cells and they may be partitioned into multiple recognizable groups, with the variability among cells within a group to be much smaller than that between groups (Kleppe and Levine, 2014). Each of such a group is referred to as a subclone. Sometimes the subclones are spatially segregated, and can be revealed by multi-regional sampling and analysis of a single tumor (Sottoriva et al., 2013, Gerlinger et al., 2012). In other cases, cells with different molecular features may be interspersed thoroughly, such that the subclonal structures are not apparent even with regions sampling down to smaller spatial units. Ultimately, single-cell profiling is the most effective strategy to study intra-tumor heterogeneity, but it incurs much higher costs in time and resources (Navin et al., 2011, Hou et al., 2012, Zong et al., 2012). Nonetheless, molecular difference among tumor subclones represents the fundamental source of drug resistance: even the most effective treatment can eradicate all subclones, and the drug-resistant cells that remain after treatment can expand and grow into the recurrent tumor. Being able to monitor clonal structure in bulk tissue samples is both an important research question and a valuable clinical capability.

1.1.3 Biological Sources of Tumor Heterogeneity

Phenotypic variability of tumor cells could be driven by multiple molecular sources of variation. One of them is genetic variation. While a person's germline genome could carry

different alleles that confer cancer susceptibility, the genome of individual cancer cells accumulates additional somatic alterations (Meyerson et al., 2010) that include copy number changes (Hanahan and Weinberg, 2000, Albertson, 2006, Beroukhim et al., 2010), translocations (Mitelman et al., 2007) and single nucleotide mutations (Sjoblom et al., 2006, Ley et al., 2008, Stratton et al., 2009). The rate of these aberrations can be increased due to impaired DNA damage repair mechanism (de Gruijl et al., 2001) or compromised surveillance of genomic instability (Negrini et al., 2010). Genetic variation is one of central components of the clonal evolution process (Nowell, 1976), which also involves natural selection, population expansion and migration, random genetic drift, interactions with local environment, and, upon effective treatment, a significant collapse of population size (Greaves and Maley, 2012). As is often the case in population genetics concerning humans or other species, only a minority of somatic mutations are capable of promoting cell survival, proliferation, and clonal expansion, and are referred to as "drivers". The majority of mutation has no discernible phenotypic impact and is referred to as "passengers" (Greenman et al., 2007).

Besides genetic variation, differences in epigenetic modifications also contribute to cancer heterogeneity (Esteller, 2008, Sharma et al., 2010). Different alterations of DNA methylation, nucleosome positioning and gene expressions among tumor cells contribute to intra-tumor heterogeneity. Such epigenetic variations also underlie the Cancer Stem Cell (CSC) Model, which emphasizes the possibility that the phenotypic variability among tumor cells can be due to epigenetic factors. Extensive studies have been conducted under this model for many cancer types (Singh et al., 2003, Prince et al., 2007, Charafe-Jauffret et al., 2009).

In addition to genetic and epigenetic heterogeneities, recent advances also highlight the interplay between tumor cells and their local tissue environment (Egeblad et al., 2010, Junttila and de Sauvage, 2013). For example, cancer associated fibroblasts promote tumor growth via secretion of multiple growth factors; active vasculature delivers necessary nutrition to proliferating tumor cells, and immune cells can be recruited and converted by tumor cells to suppress adaptive immunity and enhance tumor development. Differences in fibroblast behaviors and responses, uneven vascularization and vascular maturity, and diverse invasive immune cell types and their localizations all contribute to the observed inter- and intra-tumor heterogeneity, and must be considered when trying to overcome therapeutic resistance.

1.1.4 Applications of High-throughput Technologies in Tumor Heterogeneity Research

In recent years, the maturation of high-throughput technologies has rapidly enhanced our ability to study complex cancer genomes. For example, whole-genome sequencing of tumor samples provides a nearly complete catalog of somatic changes, including single-nucleotide variations, small insertion and deletions, and structural variations. Likewise, RNAseq technology has enabled the discovery of gene fusion (Tomlins et al., 2005, Soda et al., 2007), novel transcripts (Maher et al., 2009), and RNA editing (Sommer et al., 1991). Other high-throughput technologies, including SNP genotyping arrays, gene expression arrays, micro-RNA expression arrays, DNA methylation arrays, etc., enable the simultaneous

analyses of multiple levels of biological regulation, and have contributed to an increasingly integrative understanding of the mechanism of tumor development.

Many large-scale, coordinated studies have been conducted during the past decade to understand the etiology of cancers by applying genomic technologies. The Cancer Genome Atlas (TCGA) project and the International Cancer genome Consortium (ICGC) are two examples of such consortium-scale, highly collaborative initiatives. By the February of 2014, TCGA has analyzed more than 9,000 tumor samples for 29 types of cancers, and has publicly released both the clinical data and many types of genomic profiling data. These data resources have provided a valuable opportunity to study the molecular basis of multiple of cancer types from multiple 'omics perspectives. A limitation of these datasets, as I will discuss further in this dissertation, is that they treat each tumor sample as a unitary, homogeneous entity, and have not considered the intrinsic heterogeneity within each tumor samples. I will show in Chapter 2 that information about intra-cellular heterogeneity **can** be extracted from datasets originally intended to study the average behavior of bulk tumor samples, and intra-cellular heterogeneity can explain a substantial portion of the observed inter-tumor heterogeneity.

1.2 Challenges in studying intra-tumor heterogeneity using bulk tumor datasets

1.2.1 Normal Tissue Contamination in Inter-tumor Heterogeneity Studies

Large cancer cohorts mentioned above routinely used molecular materials collected from bulk tumor tissues that usually contain normal cells. Low tumor content reduces the power to detect somatic events using genomic DNA data, especially for subclonal mutations—those that appear in only a fraction of the tumor cells (Carter et al., 2012). Tumor-normal mixing also affects every other molecular profiles, including mRNA expression, DNA methylation, micro-RNA expression etc., by presenting a weighted average of the signatures carried by tumor cells and those carried by the normal calls. Variable ratios of tumor-normal mixing affect tumor subtype classification, thus directly complicating the clinical applications of the patients' molecular profiling data. I present a case study using Glioblastoma Multiforme samples showing this problem. GBM was one of the first cancers TCGA analyzed (TCGA, 2008). GBM is a malignant brain cancer (WHO grade IV astrocytoma). Despite intensive treatment, the outcome is poor since the median survival time is only 18 months. (Johnson and O'Neill, 2012). Researchers have studied the molecular subtypes to characterize the inter-tumor diversity of this cancer (Phillips et al., 2006a, Verhaak et al., 2010a). Verhaak et al applied the mRNA expression data from TCGA samples and discovered four subtypes. However, many TCGA studies including this one failed to consider the impact of intra-tumor heterogeneity, in Verhaak's classification, different subtypes show no significant survival difference.

In the following paragraph I am going to review the general principle on how to infer tumor/normal mixing. TCGA applied genome-wide SNP array to profile the copy number alterations in tumor samples. SNP array is an efficient technique to estimate copy number

changes and allelic imbalances both at high resolution and throughout the whole genome (Zhao et al., 2004, Dutt and Beroukhim, 2007). There are two main commercial platforms for SNP array analysis, Affymetrix and Illumina. Both platforms produce allele specific copy number estimates, initially derived from the intensity of fluorescent assay signals. For each SNP, the two alleles from a diploid genome are denoted by A and B. The intensity of both alleles (A+B) provides an estimate for the total copy number. In practice, it is convenient to use $log_2 \frac{A+B}{2} = log_2(A+B) - 1$ (logR ratio, or LRR) to represent copy number, since it takes value zero for normal diploid loci. The fraction of B allele signal intensity (B/(A+B)), normally referred to as B-allele frequency or BAF, provides evidence for allelic-imbalances. Although profiled on genomic DNA from bulk tissue, SNP array data contains intra-tumor heterogeneity information and a number of algorithms have been developed to extract tumor/normal mixing ratios using SNP array (Popova et al., 2009, Yau et al., 2010, Van Loo et al., 2010, Song et al., 2012). When the sample contains a fraction of euploid cells, in a copy number variation region, not all the cells are carrying the CNV, and instead of theoretical LRR and BAF values based on copy numbers, the observed LRR and BAF will be closer to zero, a phenomenon referred to as 'contraction'. BAF and LRR contraction is helpful to infer tumor/normal mixing ratios.

In **Chapter 2**, I re-examined Verhaak et al.'s cohort by considering intra-tumor heterogeneity. I first developed an algorithm to estimate the tumor/normal mixing ratios for individual samples using SNP array data, and discovered that the variation of mRNA expression pattern among samples is driven by different levels of euploid cell fractions of individual tumors. I then revised the classification of mRNA expression subtypes with joint use of mRNA and

CNA data. The new scheme I proposed has stronger predictive power on clinical outcome. Using inferred normal cell mixing ratios and reference datasets of known neuronal cell types, I was able to identify microglia/macrophage as the likely source of the euploid cells in the mesenchymal GBMs.

1.2.2 Tumor Subclone Analysis Using Bulk Tissues

As discussed above, intra-tumor heterogeneity is a hallmark of cancer and reflects the complicated evolution history of tumors. There are several aspects of interests of intra-tumor heterogeneity, including cellular frequencies of somatic copy number alterations (sCNAs) and somatic mutations, number of subclones in a tumor population, etc. Cells carrying somatic driver events have greater selective advantage and are likely to be maintained during tumor evolution. Therefore, somatic aberrations with high cellular frequencies are usually candidate driver events. More interestingly, if a somatic event occurred with high prevalence in a subclone, it is likely a subclonal driver event. Studying subclonal drivers is helpful to understand tumor evolution. .

The ideal data for analyzing tumor subclones is to profile single cells or samples collected from multiple regions of the same tumor. However, these procedures remain expensive and labor intensive. Bulk tissue analysis is still a common study design and generated large

amounts of data. There is therefore a strong need of analytical tools to effectively infer intra-tumor heterogeneity using such suboptimal data. In the past five years, a number of algorithms have emerged that can infer tumor subclonal features using data generated from bulk tissue. In the following I will provide a review of six methods, and discuss their advantages and limitations.

1.2.2.1 Review of Methods Studying Tumor Subclones

Carter et al (Carter et al., 2012) introduced an algorithm, ABSOLUTE to study intra-tumor heterogeneity. The segmented and smoothed copy number data is first displayed on a histogram to examine the distribution of copy ratios (normalized copy numbers). Usually these values group tightly into separate peaks on the density plot, each peak representing a copy number configuration. Due to mixing with euploid genome, the spacing between adjacent peaks (b) do not reach full theoretical values and the copy ratios (δ_τ) of regions with homozygous deletions which should be zero, are usually positive. ABSOLUTE infer euploid mixing ratios depending on b and δ_τ. A subclonal segment will generate small peak between two major peaks of clonal events, and ABSOLUTE acknowledge it as an outlier and infer tumor purity using space between major peaks. In addition to inferences made via copy number profiles, the authors also extended ABSOLUTE to estimate average allele counts per cancer cell, or cell multiplicity (s_q) using somatic mutation profiles with a Beta-Multinormial likelihood model.

Overall, ABSOLUTE made a contribution in the field by explicitly modeling subclonal events during tumor purity and ploidy estimation. Integration of somatic mutation data and copy number results is also a breakthrough. However, there were several existing challenges remained unsolved. First, ABSOLUTE lacks the capability to quantitatively estimate the cellular fractions of subclonal sCNA carriers. Second, despite the discussion on cellular multiplicity for somatic mutations, the inference is not sufficient. In their likelihood model, somatic mutations and CNAs always occur in the same lineage, while it is not always true (Nik-Zainal et al., 2012). Another limitation of this approach is that the inference of subclonal cell multiplicity remained categorical. The authors failed to provide any quantitative estimation to subclonal somatic mutations. Moreover, the inference on s_q relied on copy number determination, and since only clonal CNA events were analyzed by ABSOLUTE, for any somatic mutation, clonal or subclonal, if it hit a subclonal CNA region, no information could be concluded from the sequencing data.

Nik-Zainal et al (Nik-Zainal et al., 2012) studied the subclone structure for one breast tumor sample using whole genome sequencing (WGS) with an 188× average depth. The authors collected somatic mutations in euploid genomic regions and noticed that the distribution of somatic allele frequencies consists of four distinguishable clusters, suggesting that this tumor harbored multiple subclones. To further determine the lineage relationships between subclones, the authors phased adjacent somatic mutations that are spanned by the same read pair. If mutations from cluster X were always in phase with mutations from cluster Y, then X and Y were in the same lineage. On the other hand, if mutations from cluster X were never in phase with those from Y, X and Y belonged to different lineages. They used this approach to

determine if the cells carrying mutations from cluster X is in linear or branching relationship with cells carrying mutations from cluster Y. This approach has at least two contributions: 1) it set forward an approach to robustly identify the number of subclones in tumors using the distribution of somatic allele frequencies; 2) phasing somatic mutations provided rich information that can be used to infer lineage relationships between subclones. The major limitation is that, the analysis applied in this study is manually optimized for a few (twenty-one) samples and is not readily applied to larger cohorts.

Landau et al (Landau et al., 2013) developed an algorithm to infer the cancer cell fraction (CCF) of somatic mutations and CNAs, using both SNP array and whole exome sequencing (WES) data. For CNAs, they modified the original ABSOLUTE algorithm to model subclonal events. In Landau's method, in a subclonal region, the mixing ratio is allowed to be different from the global tumor purity, on the condition that the tumor CNAs only alter from the euploid state by one copy. This assumption, however, is unnecessary and often violated. Adding WES data, they were able to estimate the CCF of somatic mutations only in clonal CNA regions, and assuming that somatic mutation has occurred later than CNA and therefore affect only one allele. The second assumption is oversimplified and can be violated when mutation occurred before the CNA. For somatic mutations in subclonal CNA regions, they estimated CCF manually. This work made a contribution by explicitly modeling the subclonality of somatic mutations, corrected for local copy number events. . However, it is incomplete to assume that somatic mutations could only occur after CNAs. For example, if a somatic mutation occurred early in a region of subclonal copy neutral LOH, both alleles of the LOH-carrying cells would harbor the mutation. Using Landau's approach, the CCF of this

mutation would be overestimated by a factor of two, which could wrongly assign a subclonal event to be clonal.

EXPANDS (Andor et al., 2014) estimates tumor subclonal structure. It used sequencing data to infer the fraction of cells carrying a specific CNA or somatic mutation. The authors defined 'B allele' to be non-reference allele, which is different from the definition in SNP array, where B allele is arbitrarily chosen. In this definition, B allele can either be somatic mutations or germline polymorphic sites (will be AA if the site does not contain germline mutation), which is an improvement compared with previous approaches. EXPANDS estimated subclonality for each locus independently using the *in-phase* constraint, by enumeration of all the possible combinations of allele-specific copy numbers and screening of possible mixing ratios from (0,1). As mentioned above, the *in-phase* constraint is a very strong and unrealistic assumption, and using this assumption implicitly is the major limitation of this approach. Another limitation is, EXPANDS failed to consider all the possible temporal and lineage relationships between an sCNA and a somatic mutation occurred in the same locus. Due to this limitation, it cannot accurately estimate cellular frequencies of somatic mutations that have occurred in branching lineages with sCNA.

Pyclone (Roth et al., 2014) used sequencing data to estimate CCF of somatic mutations (referred to as cellular prevalence) and to perform phylogenetic analysis of tumor subclones. Pyclone relied on other methods to infer the absolute copy numbers for each locus as a prerequisite. In order to estimate CCF, Pyclone introduced five possible relationships (denoted as priors) between a somatic CNA and a mutation occurred in it. Of these five priors, predictions using the Parental Copy Number (PCN) prior were the most accurate according to

their simulations. The PCN prior considered two lineage relationships: 1) mutation occurred before CNA, but with the *in-phase* constraint; 2) mutation occurred after sCNA, but did not include the scenario when mutation and sCNA occurred in different lineages. Pyclone considered a more complete set of lineage relationships between a somatic mutation and a CNA, yet was still unsatisfactory due to the *in-phase* assumption.

The methods above estimate subclonality of somatic events independently. Another method, THeTa (Oesper et al., 2013) used an alternative approach. It jointly used all sCNAs to simultaenously estimate 1) the number of subclones in a tumor sample; 2) the abundance of each subclone and 3) the total copy number carried by each subclone in each locus. For a given number (K) of subclones, THeTa models the observed read counts (Y^i) in the genomic locus i (i=1,2…n, n is the number of loci) as the linear combination of K components: $Y^i=\mathbf{N^i}\times\boldsymbol{\mu}$, where $\mathbf{N^i}=(n_1{}^i,n_2{}^i,\ldots,n_K{}^i)$ is the copy number vector for K subclones in locus i, and $\boldsymbol{\mu}=(\mu_1,\mu_2,\ldots,\mu_K)$ is the vector of subclonal abundance. It then enumerates all the possible integer combinations of copy numbers in the K subclones and select the optimum solutions of $\boldsymbol{\mu}$. Combined usage of markers across the genome increases the reliability of inference in THeTa. However, its computational time increases exponentially with the number of markers analyzed. Also, THeTa could not infer the subclonality of somatic mutations.

To conclude, current methods developed to infer tumor subclones suffer from several limitations. First, they usually only estimate the cellular frequencies of one type of events, and none of the above quantitatively infer cellular frequencies for both sCNA and somatic mutations. Furthermore, most of these methods implicitly applied the unrealistic '*in-phase*' assumption. Finally, none of the methods above considered the scenario when a mutation

occurred in different lineage with an sCNA in the same locus. These methods cannot provide accurate estimations when their assumptions are violated.

1.2.2.2 Introduction of Clonal Heterogeneity Analysis Tool

In section ***1.2.2.2*** I have reviewed recently developed methods on intra-tumor heterogeneity. Applications of these methods provided insights into recognition of subclones (Carter et al., 2012), subclonal architecture (Nik-Zainal et al., 2012, Roth et al., 2014), dynamics of population alterations under treatment and the discovery of subclonal driver mutations (Landau et al., 2013). Despite these efforts, the field lacks a systematic tool that integrates both sCNA and somatic mutation, and provides comprehensive estimations cellular frequencies of somatic mutations and sCNAs without using limiting hypothesis such as the '*in-phase*' assumption. In **Chapter 3**, I introduce Clonal Heterogeneity Analysis Tool (*CHAT*), for inferring cellular frequencies of both sCNA and somatic mutations, by jointly analyzing DNA SNP array data and DNA sequencing data. In *CHAT*, I integrated different types of somatic events through a systematic investigation of lineage scenarios of mutations in an sCNA region. Below is a brief introduction to this topic.

For example, the task is to estimate cancer cell fraction (CCF) of a somatic mutation, which is the subclonality for mutations. Consider a somatic mutation that occurs in a euploid region and hits one chromosome: if the observed somatic allele frequency (SAF) is f, then the

estimated CCFis simply 2*f*. However, if the mutation resides in an sCNA region, the relationship between CCF and SAF depends on the copy number configuration: copy neutral loss-of-heterozygosity (CN-LOH), deletion, amplification, etc. and the cellular frequency of the sCNA. Further, it also depends on the chromosomal background in which the mutation occurs: on the parental chromosome with higher copy number (major allele) or smaller copy number (minor allele). Given the observed SAF of the mutation, it will be impossible to estimate CCF without all the information mentioned above.

A previous research (Durinck et al., 2011) studied the temporal order of somatic mutations and CN-LOH event of TP53 gene in 8 cutaneous squamous cell carcinoma samples with whole-exome sequencing data. The authors argued that when mutation occurred earlier than CN-LOH, both alleles would carry the mutation, and generate homozygous genotypes; otherwise it generates heterozygous observations. They used heterozygosity to estimate the temporal orders between the TP53 somatic mutations and CN-LOH event, and found these mutations were early events. This method modeled the temporal order of somatic events explicitly. However, it did not take into consideration the possibility that a given CN-LOH event could be subclonal, and therefore, even though the mutation occurred early, it could still appear to be heterozygous due to mixing with euploid cells. Also, the estimation is limited to CN-LOH events and not generalized to other sCNA types.

In *CHAT*, I first implemented the estimation of sCNA genotypes and sCNA subclonality. Then, to infer cellular frequencies of somatic mutations, I considered the following scenarios:

A) The mutation and sCNA emerged sequentially, with the mutation occurring first, and the

sCNA occurring in a subset of mutation-bearing cells. Cells carrying both mutation and sCNA may have two configurations: $\mathbf{A_1}$: the duplication occurred on the mutation-bearing chromosome, and $\mathbf{A_2}$: the duplication occurred on the mutation-free chromosome.

B) Like **A**, the mutation and sCNA emerged sequentially; but unlike **A**, the sCNA occurred first, with the mutation occurring in a subset of sCNA-bearing cells. Mutation may have occurred on one of the duplicated chromosome ($\mathbf{B_1}$) or the un-duplicated chromosome ($\mathbf{B_2}$).

C) The mutation and sCNA emerged independently, i.e., appearing in non-overlapping populations of cells.

All previous methods only considered a subset of these scenarios. For example, Landau et al.'s approach only considered scenario **B**, while EXPANDS considered scenarios $\mathbf{A_1}$ and $\mathbf{A_2}$, and PyClone considered $\mathbf{A_1}$, $\mathbf{A_2}$ and **B**, but failed to include **C**. With systematic investigation of lineage scenarios, CHAT is able to estimate cell fractions for somatic mutations without limiting assumptions.

1.2.3 Detection and Genotyping Short Tandem Repeats in Complex Genomes

Previous cancer research extensively studied somatic CNAs and mutations, while other types of genomic aberrations remain poorly understood, including many kinds of structural variations. There is need in the field to understand the role of short tandem repeats (STR), or microsatellites in human diseases, including cancer. Short tandem repeats are consecutive

occurrence of 2-6 bases of DNA sequence many times. STR locI am very common in the human genome (Willems et al, 2014) and analysis of STR is useful in many fields, including forensic usage, paternity test (Jobling et al., 1997), phylogenetic analysis (Jarne and Lagoda, 1996), etc. For example, germline mutations of STR locI am responsible for many neurodevelopmental disorders, including Huntington's disease (Walker, 2007), Fragile X syndrome (Pearson et al., 2005) and multiple types of spinocerebellum ataxia (Pulst et al., 1996, Campuzano et al., 1996, Paulson, 2012). Somatic changes in STR length are common in some cancers, in a phenomenon known as microsatellite instability (MSI). MSI is caused by impaired DNA mismatch repair (Liu et al., 1995, Ellegren, 2004), and it has been characterized in colorectal cancer (Popat et al., 2005) and prostate cancer (Uchida et al., 1995). Current methods to genotype STR loci remain slow and labor-intensive. Traditional Sanger sequencing technology still serves as a gold standard in determining the number of repeats, yet cannot be efficiently applied to large sample cohorts or to genome-wide analyses. The development of next-generation sequencing technologies allows researchers to analyze many samples in genome-wide scale, but to genotype STR alleles using short-read sequencing data is a novel challenge because many STR alleles are longer than the read length. Two algorithms, lobSTR (Gymrek et al., 2012) and RepeatSeq (Highnam et al., 2013) have been developed to address part of this challenge. However, both methods are limited to genotype STR alleles that are shorter than read length, which is typically 100-nt for the Illumina HiSeq2000 sequencer. These methods are not suitable to detect abnormally expanded STR alleles beyond the read length, nor can they discover novel microsatellite regions due to their reliance of the locations of known STR loci. In **Chapter 4**, I introduce a

new algorithm, called *STRfinder*, using paired-end short read sequencing data to genotype STR loci. *STRfinder* can detect novel STR loci and genotype STR alleles that are much longer than a read, and out-performs lobSTR or RepeatSeq for these alleles in terms of variant call rates, genotyping accuracy, and length estimation precision.

1.3 Summary

In this Chapter, I reviewed the concepts and background related to tumor heterogeneity. Large amount of cancer -omics data have been accumulated in recent years, and most of these involve one-sample-per-tumor, bulk tissue analysis. A number of methods have been developed to analyze these data to infer the features of intra-tumor heterogeneity. However, there are at least three challenges in the field. First, studies of inter-tumor heterogeneity among multiple samples usually failed to consider intra-tumor heterogeneity, and the results could be confounded by tumor/normal mixing ratios or tumor subclones. In **Chapter 2**, I have addressed this challenge. Second, analytical tools specifically developed for analyzing intra-tumor heterogeneity suffer from limiting assumptions reviewed in previous sections, and *CHAT* overcomes these limitations. Third, as an important class of genetic variation that can underlie the risks of both constitutional and somatic diseases, short tandem repeat has not received enough attention. There is need to detect and genotype STR alleles both in large sample cohorts and genomewide, using next generation sequencing data. In my thesis, I aimed to address the above challenges by developing novel bioinformatics tools.

1.4 Bibliography

Albertson, D. G. 2006. Gene amplification in cancer. *Trends Genet,* 22**,** 447-55.

Andor, N., Harness, J. V., Muller, S., Mewes, H. W. & Petritsch, C. 2014. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics,* 30**,** 50-60.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M.*, et al.* 2010. The landscape of somatic copy-number alteration across human cancers. *Nature,* 463**,** 899-905.

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A.*, et al.* 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science,* 271**,** 1423-7.

Carter, S. L., Cibulskis, K., Helman, E., Mckenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A.*, et al.* 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol,* 30**,** 413-21.

Charafe-Jauffret, E., Ginestier, C., Iovino, F., Wicinski, J., Cervera, N., Finetti, P., Hur, M. H., Diebel, M. E., Monville, F., Dutcher, J.*, et al.* 2009. Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res,* 69**,** 1302-13.

De Gruijl, F. R., Van Kranen, H. J. & Mullenders, L. H. 2001. UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer. *J Photochem Photobiol B,* 63**,** 19-27.

Durinck, S., Ho, C., Wang, N. J., Liao, W., Jakkula, L. R., Collisson, E. A., Pons, J., Chan, S. W., Lam, E. T., Chu, C.*, et al.* 2011. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov,* 1**,** 137-43.

Dutt, A. & Beroukhim, R. 2007. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol,* 19**,** 43-9.

Egeblad, M., Nakasone, E. S. & Werb, Z. 2010. Tumors as organs: complex tissues that interface with the entire organism. *Dev Cell,* 18**,** 884-901.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet,* 5**,** 435-45.

Esteller, M. 2008. Epigenetics in cancer. *N Engl J Med,* 358**,** 1148-59.

Greaves, M. & Maley, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C.*, et al.* 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-8.

Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res,* 22**,** 1154-62.

Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57-70.

Heppner, G. H. 1984. Tumor heterogeneity. *Cancer Res,* 44**,** 2259-65.

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A. & Mittelman, D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res,* 41**,** e32.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D.*, et al.* 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative

myeloproliferative neoplasm. *Cell,* 148**,** 873-85.
Jarne, P. & Lagoda, P. J. 1996. Microsatellites, from molecules to populations and back. *Trends Ecol Evol,* 11**,** 424-9.
Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T. & Thun, M. J. 2008. Cancer statistics, 2008. *CA Cancer J Clin,* 58**,** 71-96.
Jobling, M. A., Pandya, A. & Tyler-Smith, C. 1997. The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med,* 110**,** 118-24.
Johnson, D. R. & O'neill, B. P. 2012. Glioblastoma survival in the United States before and during the temozolomide era. *J Neurooncol,* 107**,** 359-64.
Junttila, M. R. & De Sauvage, F. J. 2013. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature,* 501**,** 346-54.
Kleppe, M. & Levine, R. L. 2014. Tumor heterogeneity confounds and illuminates: assessing the implications. *Nat Med,* 20**,** 342-4.
Landau, D. A., Carter, S. L., Stojanov, P., Mckenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L.*, et al.* 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell,* 152**,** 714-26.
Levan, A. & Hauschka, T. S. 1953. Endomitotic reduplication mechanisms in ascites tumors of the mouse. *J Natl Cancer Inst,* 14**,** 1-43.
Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., Mclellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., Mcgrath, S., Hickenbotham, M.*, et al.* 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature,* 456**,** 66-72.
Liu, B., Nicolaides, N. C., Markowitz, S., Willson, J. K., Parsons, R. E., Jen, J., Papadopolous, N., Peltomaki, P., De La Chapelle, A., Hamilton, S. R.*, et al.* 1995. Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nat Genet,* 9**,** 48-55.
Maher, C. A., Palanisamy, N., Brenner, J. C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T. R., Grasso, C., Yu, J.*, et al.* 2009. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A,* 106**,** 12353-8.
Marte, B. 2013. Tumour heterogeneity. *Nature,* 501**,** 327.
Marusyk, A., Almendro, V. & Polyak, K. 2012. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer,* 12**,** 323-34.
Marusyk, A. & Polyak, K. 2010. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta,* 1805**,** 105-17.
Meyerson, M., Gabriel, S. & Getz, G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet,* 11**,** 685-96.
Mitelman, F., Johansson, B. & Mertens, F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer,* 7**,** 233-45.
Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D.*, et al.* 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.
Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. 2010. Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol,* 11**,** 220-8.
Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* 2012. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

Nowell, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

Oesper, L., Mahmoody, A. & Raphael, B. J. 2013. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol,* 14**,** R80.

Paulson, H. 2012. Machado-Joseph disease/spinocerebellar ataxia type 3. *Handb Clin Neurol,* 103**,** 437-49.

Pearson, C. E., Nichol Edamura, K. & Cleary, J. D. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet,* 6**,** 729-42.

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L.*, et al.* 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell,* 9**,** 157-73.

Popat, S., Hubner, R. & Houlston, R. S. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol,* 23**,** 609-18.

Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E. & Stern, M. H. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol,* 10**,** R128.

Prince, M. E., Sivanandan, R., Kaczorowski, A., Wolf, G. T., Kaplan, M. J., Dalerba, P., Weissman, I. L., Clarke, M. F. & Ailles, L. E. 2007. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proc Natl Acad Sci U S A,* 104**,** 973-8.

Pulst, S. M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X. N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A.*, et al.* 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet,* 14**,** 269-76.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Cote, A. & Shah, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods,* 11**,** 396-8.

Sharma, S., Kelly, T. K. & Jones, P. A. 2010. Epigenetics in cancer. *Carcinogenesis,* 31**,** 27-36.

Singh, S. K., Clarke, I. D., Terasaki, M., Bonn, V. E., Hawkins, C., Squire, J. & Dirks, P. B. 2003. Identification of a cancer stem cell in human brain tumors. *Cancer Res,* 63**,** 5821-8.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N.*, et al.* 2006. The consensus coding sequences of human breast and colorectal cancers. *Science,* 314**,** 268-74.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H.*, et al.* 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature,* 448**,** 561-6.

Sommer, B., Kohler, M., Sprengel, R. & Seeburg, P. H. 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell,* 67**,** 11-9.

Song, S., Nones, K., Miller, D., Harliwong, I., Kassahn, K. S., Pinese, M., Pajic, M., Gill, A. J., Johns, A. L., Anderson, M.*, et al.* 2012. qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One,* 7**,** e45835.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C. & Tavare, S. 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A,* 110**,** 4009-14.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-24.

Tcga 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455**,** 1061-8.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R.*, et al.* 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science,* 310**,** 644-8.

Uchida, T., Wada, C., Wang, C., Ishida, H., Egawa, S., Yokoyama, E., Ohtani, H. & Koshiba, K. 1995. Microsatellite instability in prostate cancer. *Oncogene,* 10**,** 1019-22.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B.*, et al.* 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P.*, et al.* 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* 17**,** 98-110.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546-58.

Walker, F. O. 2007. Huntington's Disease. *Semin Neurol,* 27**,** 143-50.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A.*, et al.* 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467**,** 1114-7.

Yates, L. R. & Campbell, P. J. 2012. Evolution of the cancer genome. *Nat Rev Genet,* 13**,** 795-806.

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O. & Holmes, C. C. 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol,* 11**,** R92.

Zhao, X., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T. H., Girard, L., Minna, J., Christiani, D., Leo, C.*, et al.* 2004. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res,* 64**,** 3060-71.

Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science,* 338**,** 1622-6.

Chapter 2. Inference of aneuploidy genome proportion and revised classification of human glioblastoma multiforme (GBM)

2.1 Introduction

Glioblastoma Multiforme (GBM) is an aggressive brain tumor with poor prognosis (Adamson et al., 2009). Recently, genomic profiling studies have provided rich new information for understanding molecular lesions in GBM. For example, the Cancer Genome Atlas (TCGA) project characterized several hundred GBM samples, of which many were analyzed across multiple dimensions, including single nucleotide polymorphism (SNP) genotyping, mRNA and microRNA (miRNA) profiling, DNA sequencing, and promoter methylation analysis (The Cancer Genome Atlas Research Network, 2008). These data highlighted the importance of *ERBB2*, *NF1* and *TP53* genes, and revealed recurrent aberrations in the *RTK/RAS/PI(3)K*, *p53*, and *RB* signaling pathways. Meanwhile, genomewide datasets are also useful for characterizing biological diversity in a tumor collection, as evidenced by numerous reports of molecular subtypes for many cancers based on gene expression cluster analyses (Alizadeh et al., 2000, Perou et al., 2000). In particular, gene expression data for TCGA's first GBM cohort were reported to reveal four subclasses (Verhaak et al., 2010b): Proneural (PN), Neural (NL), Classical (CL) and Mesenchymal

(MES).

However, while the availability of multiple data types in TCGA provides the opportunity for combined analyses, the four-class model was based solely on mRNA expression data. DNA copy number alteration (CNA) patterns were summarized *post hoc*, not incorporated in the initial class discovery. Methylation data were analyzed subsequently (Noushmehr et al., 2010), and revealed three clusters, which lacked a clear correspondence with the four transcriptome-based classes. Furthermore, the relationship of the four-class model with those previously reported for independent datasets (Murat et al., 2008, Phillips et al., 2006b, Sun et al., 2006) was not clarified. While the differences *between* studies could be explained by variations in sample selection criteria, experimental platforms, and analysis methods, the discrepancies among different data types *within* the TCGA's collection remained un-reconciled. My first goal was therefore to combine the CNA and expression data to provide a more integrated view of the molecular diversity in GBM.

My second goal was to study *within-tumor* heterogeneity. Surgically obtained solid tumor samples (GBM included) often contain both aneuploid cells and euploid cells. I developed a method to leverage the allele-specific CNA data to estimate the fraction of aneuploid cells in each sample, and to incorporate this measure of tumor "purity" in class discovery. I also asked if results in GBM were seen in the ovarian (OV) cancer cohort from TCGA (The Cancer Genome Atlas Research Network, 2011). I emphasized between-cohort concordance in deciding the optimal number of clusters, and annotated the potential cell type of origin of different classes by comparing GBM gene expression data to reference datasets of known cell types. My results led to a revised framework of GBM classification, and I sought to

understand its biological implication and clinical relevance. I validated the between-class difference in survival time in an independent GBM cohort. Finally, I summarized the newly recognized subclasses and associated biomarkers into a hierarchical classification protocol for use in diagnosis and further research.

2.2 Data sources

2.2.1 Glioblastoma Multiforme (GBM)

This study covered three cohorts of GBM samples. **GBM1** is the cohort analyzed by the TCGA pilot study (The Cancer Genome Atlas Research Network, 2008, Verhaak et al., 2010b). A second cohort was subsequently available and was called **GBM2** (Verhaak et al., 2010b). For validating the survival time differences I selected additional samples that became available by early 2012, and called it **GBM3**.

2.2.1.1 DNA copy number data

For GBM1-2, Allele-specific copy number data for Illumina HumanHap550K arrays were downloaded from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/) on 4/14/2010. I queried the Data Access Matrix by choosing

Disease: *GBM*;

Data Type: *SNP*;

Data Level: *2 and 3*;

Platform: *HAIB (HumanHap550)*.

This query yielded tumor-normal logR Ratio (LRR) data for 284 paired samples, and B allele frequency (BAF) data for 347 tumor samples, of which 284 had matched normal samples. The overlapping set of 284 paired samples, 130 in GBM1 and 154 in GBM2, was selected for further analysis. The dataset contains 561,468 autosomal SNPs.

Allele-specific copy number data for GBM3 came from TCGA batches 26, 38, 62, 79, 111, and 130. A total of 156 samples had both Affymetrix SNP 6.0 genotyping data and Affymetrix gene expression data available, and these were downloaded in bulk on 1/31/2012. The copy number data for Affymetrix SNP arrays were Birdsuite output files and were converted to logR and BAF. Ten samples were apparent outliers on the gene expression PCA plot (not shown), and were removed. Of the remaining 146 samples, two female patients (TCGA-12-3644 and TCGA-12-3646) had exceptionally longer survival time (62 and 44 months respectively), and were removed before classification analysis and survival time comparisons. The genome coordinates for the 811 autosomal cytoband were from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBand.txt.gz.

2.2.1.2 Gene expression data

Gene expression data were downloaded from

http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/. Most of our analyses were based on "unifiedScaledFiltered.txt", which contains processed data for 1,740 most variable genes for 202 GBM samples. The data processing procedure was reported previously (Verhaak et al, 2010). I also analyzed the full dataset in "unifiedScaled.txt", containing 11,861 genes before filtering. In GBM1, a subset of 130 samples (out of 202) had both gene expression and DNA copy number data. In GBM2, all 154 samples had both gene expression and DNA copy number data.

Expression data for GBM3 were downloaded on 1/9/2012, from TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm), by choosing Platform= BI_HT_HG-U133A. GBM3 contains 170 samples from TCGA batches 26, 38, 62, 19 and 111. I quantile normalized these data before running downstream analysis. The 840 genes selected in Verhaak et al. for distinguishing the four former subtypes were provided at (http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/ClaNC840_centroids.xls) and accessed on 04/12/2011.

2.2.1.3 MicroRNA data

MicroRNA data were downloaded on 1/7/2012 from TCGA portal by choosing

Data Type: Expression miRNA;

Batch: *All*;

Level: *3;*

Platform: *UNC_miRNA_8×15K*,

This returned a dataset for 534 miRNAs in 506 samples, of which 125 overlapped with the 202 GBM1 tumors. I quantile normalized these data before running downstream analysis.

2.2.1.4 Clinical information

Clinical data for individual patients and samples were downloaded from TCGA data-access site:

http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/bcr/intgen.org/minbiotab/clin/ on 1/18/2011. I extracted information regarding age of diagnosis, survival time, tumor cell, and tumor nuclei. An updated version, containing information for GBM3, was accessed on 1/5/2012. The Karnofsky performance scores were extracted from the clinical data accessed on 6/22/2012.

2.2.2 Ovarian Cancer (OV)

2.2.2.1 DNA copy number data

Copy number data for Illumina 1M-Duo arrays were downloaded from TCGA Bulk Download site http://tcga-data.nci.nih.gov/tcga/findArchives.htm on 10/01/2010. I queried the Data Access Matrix by choosing

Disease: *Ovarian Cancer*;

Data Type: *SNP*;

Batch number: *all*;

Data Level: *2 and 3*;

Platform: *HAIB (Human1MDuo)*.

The BAF files contain 1,199,189 SNP markers and 516 paired samples. The LRR files contained 530 samples, of which 509 were paired. The overlapping set of 509 paired samples was selected for further analysis.

2.2.2.2 Gene expression data

Gene expression data for OV were obtained from TCGA data analysis working group. The file "TCGA_batch9-15_17-19_21-22_24.UE.txt" dated 05/05/2010 contains 11,864 genes and 524 samples, of which 504 overlap with the DNA copy number dataset.

2.2.3 Phillips et al. dataset

Phillips et al. data were accessed from GEO dataset GSE4271. The processed gene expression data for 56 samples were obtained at http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/.

2.2.4 Cahoy et al. dataset

Cahoy et al. data were accessed from GEO dataset GSE9956.

2.2.5 Data for microglia/macrophage

I queried of Gene Expression Omnibus dataset (http://www.ncbi.nlm.nih.gov/gds) (Edgar et al., 2002b) to identify gene expression profiles for microglia/macrophage cells. During January 5-7th, 2012 I searched for keywords "microglia" AND "human" and found 23 independent datasets. Among these, I selected experiments for tumor cells, and this resulted in 2 datasets, GSE25289 and GSE16119, which I used to infer the likely cell types contributing to the euploid population in MES tumors.

2.3 Inferring aneuploidy genome proportion

2.3.1 Introduction to SNP array data

Allelic intensity data from SNP genotyping arrays provide quantitative copy number information of the two parental chromosomes: n_A and n_B. In a homogeneous cell population n_A and n_B are both integers, such that the logarithm of total intensity, $\log R=\log(n_A+n_B)$, and the observed B allele frequency, $BAF=n_B/(n_A+n_B)$, adopt a finite combination of discrete values, which can be shown as "canonical positions" in the BAF-LRR plot (**Figure 2.1A**). In a tumor sample, however, the population of aneuploid cells may be mixed with euploid cells, consequently logR and BAF of the former "contract" towards those of the latter; and different mixing ratios result in different degrees of contraction (**Figure 2.1B**). An example of such a

mixed GBM sample is shown in **Figure 2.2A**. Based on this feature I developed an algorithm to quantitatively estimate genomewide mixing ratio from SNP data. In the following sections 3.3-3.9, I will outline the procedures of this algorithm, including data preprocessing, theoretical models, inference and validation.

2.3.2 Two-way mixing model and aneuploidy genome proportion (AGP)

Before introducing the details of my algorithm, it is important to layout concepts and hypothesis. In this study I define *Aneuploid Content*, or synonymously, *Aneuploid Genomic Proportion (AGP*), as the parameter p in a mixture model consisting of two homogeneous populations: (1) aneuploid cells, at the fraction of p, and (2) euploid cells, at (1-p). This model has been routinely used in the field and hereby referred to as the two-way mixing model. Euploid cells carry a balanced set of parental chromosomes representing full-integer multiples of the haploid genome, and may include normal stromal cells surrounding the tumors as well as tumor cells without apparent genomic aberrations (e.g., only point mutations). Aneuploid cells, in contrast, carry CNAs at some chromosomes or subchromosomal intervals, resulting in an unbalanced set of genomic segments, each of which still contain an integer combination of parental DNA, e.g., n_A=2, and n_B=1 in a region of amplification. For many tumors, the two-way mixing model considered here is likely an over-simplification, as multiple subpopulations of tumor cells may exist, each carrying a different integer combination of parental segments. However, a mixture model with three or

more subpopulations is computationally intractable using the observed averages of the entire population; and realistically, many tumors may contain a dominant aneuploid population. A two-way mixing model is the simplest scenario that could have generated the observed data regarding varying levels of contraction in different samples. I therefore applied this model for the first-order estimation of within-tumor heterogeneity.

2.3.3 Data processing, DNA segmentation, and merging

Throughout this chapter, I focused on somatic events, defined by the differences between tumor-normal pairs, thus ignoring inherited aneuploidy.

Seven GBM samples, TCGA-06-0139, TCGA-06-0160, TCGA-06-0165, TCGA-06-0167, TCGA-06-0189, TCGA-06-0240, and TCGA-06-0881 bear few copy number alterations (CNAs), and were excluded from further analysis.

As homozygous locus is uninformative for detecting changes in BAF patterns I focused on BAF data at heterozygous loci. For each tumor-normal pair, loci with BAF value >=0.9 or <=0.1 in the normal sample were designated as homozygous. Altering the stringency of this definition did not make a major impact on AGP inference, as AGP will be driven by large aneuploid events, for which having more or fewer heterozygous loci would not substantially change the estimate of "contraction" (see below). For the heterozygous loci thus defined, I extracted tumor BAF data and generated the "folded BAF", defined as the absolute value of (BAF-0.5), for segmentation. For both GBM and OV, I performed segmentation on folded

BAF using the Circular Binary Segmentation (CBS) algorithm, implemented in the R package *DNAcopy* with default parameters, except that "minimal markers required" was set to 5. The series of BAF change points were merged with the corresponding LRR change points, which were generated by Dr. Devin Absher at the HudsonAlpha Institute of Biotechnology using CBS (Olshen et al., 2004), and were made publicly accessible as TCGA Level 3 data. As the BAF segments and LRR segments sometimes captured the same event, I merged the combined change points as follows: if a BAF change point was within 5 markers of a LRR change point, either upstream or downstream, it was removed, i.e. only the LRR breakpoint was kept, under the assumption that the two change points captured the same event, but the BAF change point was less accurately placed due to the constraint of using only heterozygous markers. After merging, small segments, defined as containing fewer than 10 BAF markers were merged with adjacent segments by removing the flanking change points. These steps resulted in a final set of CNA segments for each tumor sample.

2.3.4 Per-segment summary of LRR and BAF

For each segment in the final CNA call set, I re-calculated the median LRR and mean folded BAF to update these values within each merged segment. For segments with balanced parental chromosomes, BAF values at heterozygous locI am distributed as one track near 0.5, but it may not be centered exactly at 0.5. Likewise for segments with unbalanced parental chromosomes, BAF values split to two tracks, which may not fall symmetrically around 0.5.

To increase the accuracy of BAF estimation I fit each segment's distribution of heterozygous BAF values as either one Gaussian distribution or the summation of two Gaussian distributions. When there were in fact two tracks but the separation between tracks was small, the summed distribution might resemble a single Gaussian distribution. I used the baseline variance of BAF as the criterion to distinguish the two cases: segments with folded BAF standard variation >=0.1 were considered as two-track segments, and fit with two Gaussians. For segments with one track, I obtained the best fitting distribution as $N(\mu,\sigma)$, and defined the folded BAF value as 0. For segments with two tracks, the best fitting distribution is $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$, and the folded BAF value is $|\mu_1 - \mu_2|/2$. If the distribution cannot be fit in R or if the segment had fewer than 100 markers, the folded BAF value is taken as the mean absolute deviation around the mean: $mean(|x_i - mean(x_i)|)$, where x_i is the BAF value at the i-th marker.

2.3.5 LRR scale-normalization

The primary goal of Illumina's data processing algorithm is to find clusters that represent discrete genotypes. As a result the LRR values are not linearly scaled with the true copy number changes, e.g., when the true DNA copy number drops from 2 to 1, LRR drops less than 1 unit (2-fold). Moreover, the severity of this "saturation effect" is different between amplifications and deletions. For Illumina 550K arrays, the correction factors are 0.572 for deletions and 0.553 for amplifications (Peiffer et al., 2006). I re-scaled LRR segmental means by these ratios before downstream analysis.

2.3.6 BAF-LRR plot: canonical points and tracks

I used the *BAF-LRR plot* to depict the bivariate data of allele-specific copy numbers. In this plot, the folded BAF values are shown on the x-axis, and the normalized LRR values are shown on the y-axis. Each segment is plotted as a point in the BAF-LRR space, with the symbol size indicating segment length. Amplifications, deletions, and copy-neutral LOH segments are uniquely placed in the plot.

Canonical points, representing integer combinations of A and B alleles, were placed as follows. For a pair of integers (n_B, n_T), where n_T denotes the copy number for the B allele, and n_T denotes the total copy number of both alleles, its x and y coordinates are:

$$x = \left| \frac{n_B}{n_T} - 0.5 \right|$$

$$y = \log_2 \frac{n_T}{2} - y^0_{pl} \qquad (1)$$

where y^0_{pl} is an adjustable offset of LRR level to reflect (1) the average ploidy of the aneuploid population, which can be a non-integer, and (2) potential alternative ploidy of the euploid population. In some tumors, the euploid portion might be n_T=4 (or n_T=6) rather than n_T=2, yet the normalization procedure of each sample tended to center its genomic average LRR to 0, thus an offset is needed to adjust the y-positions of the canonical points to achieve a maximal fit. I will separately fit n_T=2, 4, and 6 for the euploid population when searching for the optimal AGP (see below).

Tumor samples that contain a mixture of euploid cells and aneuploid cells will show a *contraction* of canonical points from its original position toward the origin, where the euploid segments reside. The paths of the contraction when AGP decreases from 1 to 0 are called *canonical tracks* (**Figure 2.1A** and **2.1B**). For a given p, canonical points on the BAF-LRR plot can be organized into a 2D lattice, in which the near-vertical gridlines connect points of equal n_B. The first line, located at the right, contains all LOH points with (n_B, n_T) =(0,1), (0,2), (0,3), etc. The second line, to the left, contains (n_B, n_T) =(1,2), (1,3), (1,4), etc. And the third line contains (n_B, n_T) = (2,4), (2,5), (2,6),etc. They are orthogonal to the canonical tracks, which describe the movement of canonical points toward the origin ((n_B, n_T) = (1,2)) under shrinking values of AGP. The relative positions of the canonical points contain information for distinguishing the alternative ploidy of the euploid genome, which define the origin of contraction for the aneuploid segments.

2.3.7 Inference of Aneuploid Genome Proportion

A. Definition of Euploid Segments: On a BAF-LRR plot, euploid regions tend to land near the point (x,y) = (0,0). But due to random noise and various technical artifacts some segments may lie slightly off (0,0). Precise assignment of the near (0,0) segments into the euploid cluster is important because it affects the relative distances to other canonical points and the AGP estimates. To anchor its position, I first ran k-means clustering 10 times on the observed BAF-LRR values for all segments. For each run, I identified the segments that belong to the

cluster nearest to (0,0), and tagged them as euploid. Segments that were tagged more than 6 times out of 10 were used to define the seed position, located at the cluster mean (x_s, y_s) of the tagged segments, weighted by segment size. Second, I examined each non-seed segment to see if its coordinates (x, y) were sufficiently close to the seed location. If $|x - x_s| <= \sigma_{BAF}$ and $|y - y_s| <= \sigma_{LRR}$, this segment was "pulled" into the euploid cluster, where $\sigma_{BAF} = 0.04$, the empirically estimated standard variation of BAF, and $\sigma_{LRR} = 0.16$, the empirical standard variation of LRR. This step was iterated, with more segments joining the euploid cluster until the cluster was no longer updated. The final coordinate of the weighted center of the euploid cluster is denoted as (x_0^f, y_0^f).

B. Canonical Points under admixture: Consider the mixture containing a population of cells carrying an aneuploid segment (n_B, n_T), and a second population of cells carrying an euploid segment (n_{pl}, $2n_{pl}$), n_{pl}=1,2, or 3, and that the euploid portion makes up 1-p of the total (i.e., AGP = p). The coordinates for the mixed population are given by:

$$x = \left| \frac{p \times n_B + (1-p) \times n_{pl}}{p \times n_T + 2 \times (1-p) \times n_{pl}} - 0.5 \right| + x_0^f$$
$$y = \log_2 \frac{p \times n_T + 2 \times n_{pl} \times (1-p)}{2} - y_{pl}^0 + y_0^f \qquad (2)$$

C. Aneuploid genome Proportion: For each sample, after the euploid cluster was defined, I searched for the best fitting p and n_{pl} by screening the parameter space of p ∈ (0,1), and n_{pl} ∈ (1,2,3,4). I did not include the canonical point for homozygous deletions because their BAF or LRR values are not determined.

For each (p, n_{pl}) combination being considered, the canonical points were calculated and the penalized sum of squared distance ($SSD'_{pl}(p)$) was calculated as:

$$SSD'_{pl}(p) = \sum_{i \in \Omega} d^{i}_{\min, pl}(p) + D_{pl}(p) \quad (3)$$

where i is the segment index and Ω represents all segments in this sample (excluding those in the euploid cluster), pl stands for ploidy n_{pl}, $d^{i}_{\min, pl}$ is the squared distance of the segment to the nearest canonical point. $D_{pl}(p)$ is the penalty score for applying a larger n_{pl}, as increasing n_{pl} results in a larger number of available canonical points to fit with, and consequently a smaller sum of squared distances. Applying this penalty will avoid making excessively high euploid baseline assignments. $D_{pl}(p)$ is linearly correlated with the approximate distance between adjacent canonical points such as (2.4) and (1,4). I defined

$D_{pl}(p)$ = $p_n \times (n_{pl}-1) \times$ distance between canonical points (2,4) and (1,4)

Penalty P_n was manually chosen as p_n=200 as it generated the most reasonable assignments. Best fitting AGP value was determined by the smallest $SSD'_{pl}(p)$. The scanning of the parameters was carried out in two steps to increase computation speed: a coarse scan of p ∈ (0.05,0.95) at an interval of 0.05 was performed, with a best fitting value p* determined. Then, in the second step, a finer scan of p ∈ (p*-0.1,p*+0.1) at an interval of 0.02 was performed to refine the final score. The model also yielded the optimal ploidy value, resulting in 135 diploid, 127 tetraploid and 22 hexaploid samples for GBM1 and GBM2. For OV, I identified 23 diploid, 64 triploid, 273 tetraploid, 127 hexaploid and 22 octoploid samples.

2.3.8 Genomic features and QC measures

I extracted multiple genomic measures for each tumor sample, including percent of genome

changed (PC) and percent of genome on canonical points (PoP). Let P_1 denote the proportion of genome in the euploid cluster, thus 1-P_1 of genome has been alerted either in copy number or in the B allele frequency. I define:

$$PC = 1 - P_1$$

Extremely low PC indicates that there is insufficient amount of CNAs to inform model parameters and should be considered as having yielded low-quality AGP estimates.

Actual segments on the BAF-LRR plot may fall near or far from a canonical point for a given AGP. I quantify these deviations as measures of goodness-of-fit by the two-way mixing model with optimal AGP. If there is more than one dominant aneuploid population mixed with the euploid population, some segments would have a different mixing ratio than some other segments, and as a consequence, the fit at a single AGP would not be suitable for all segments, and this can be reflected by a low rate of "Percent-on-Point", defined as the proportion of segments falling within $s_{BAF} = 0.04$ and $s_{LRR} = 0.16$ of a canonical point. If this proportion is P_2 of the genome, I define

$$PoP = \frac{P_2}{1 - P_1}$$

As aneuploid cells carry variable copy numbers at different segments, it is no longer sufficient to define an integer ploidy as a genomewide attribute of a tumor. However I define *average aneuploid ploidy* as the genomewide mean copy number for the aneuploid cells of the tumor, and *average overall ploidy* as the weighted average of euploid and aneuploid populations. Specifically, as I assign ploidy status for every segment in the aneuploid genome, the *average aneuploid ploidy*, Ψ_{tumor}, can be defined as the length-weighted means of segmental ploidy. The *average overall ploidy* of the sample, containing p of aneuploid

genome and 1-p of euploid genome, is

$$\Psi_{overall} = \Psi_{tumor} \times p + 2 \times n_{pl} \times (1 - p)$$

Other tumor genomic features, including percentage of genome amplified (%amp), deleted (%del), percentage of hemizygous deletion (%del.loh), and percent of genome underwent loss of heterozygosity (%LOH), were also extracted.

I use a bootstrap method to estimate the confidence intervals of AGP. A weighted resampling was performed for each sample, such that each segment was chosen with the probability proportional to its size. Permutation was done 100 times for each sample, and for each run, 80% segments were resampled and AGP recalculated. The standard deviation, and the 2.5%, 50%, and 97.5% quantiles of AGP, were extracted and included in **Table 2.1**. The 2.5-97.5% *confidence interval* (CI) can be calculated from these results. I also calculated the relative confidence interval (rCI) as the ratio of CI to the median of AGP. Eighty-eight percent of samples had rCI less than 100%.

PoPs were negatively correlated with AGPs (**Figure 2.1C,** Spearman's r = -0.40, P = 3,3 × 10^{-12}), suggesting that samples better accounted for by the model (i.e., higher PoP) tend to have lower AGP estimates, thus our method may have over-estimated AGP for poorly fit samples. The CIs, however, were positively correlated with AGPs (**Figure 2.1D,** r = 0.13, P = 0.03).

2.3.9 Validation of AGP algorithm

The validation dataset, GSE11976, was downloaded from the Gene Expression Omnibus (GEO) (Edgar et al., 2002a). It contained 11 samples of DNA from the human breast carcinoma cell line CRL2324 mixed with DNA from the lymphoblastoid cell line HCC1395BL with known mixing ratios. Samples were measured across 370,404 SNP loci by using the Illumina HumanCNV370-Duov1 BeadChips. Known CNVs in HCC1395BL were removed so that HCC1395BL DNA represents the euploid portion of the mixture. AGP value for each sample was calculated using our algorithm, and compared with the mixing percentage. Pearson's correlation coefficient r = 0.979, confirming that our method accurately estimated aneuploid content.

2.4 GBM samples AGP estimation

As mentioned above, in the first batch (GBM1), seven of 284 tumors had too few CNAs (including copy-neutral loss-of-heterozygosity events) for AGP estimation, and were removed. The remaining 277 tumors had > 0.5% of the genome affected by CNAs, with an average Percent Changed (PC) of 37.3%, i.e., > 1/3 of the genome was altered in an "average" GBM. Across the 277 samples, the estimated AGPs ranged from 23% to 99% (mean ± SD: 76% ±17%), indicating significant admixture of euploid cells (average euploid content of 24%). To assess the goodness-of-fit for each sample I quantified the confidence interval (CI, 2.5-97.5%) of AGP and the fraction of CNAs that fall on canonical positions (PoP, Percent-on-Point) in the optimal two-way mixing model (**Figure 2.1C-D)**. PoP values

had a median of 92% among 277 GBMs, suggesting that it is indeed adequate to model a single dominant aneuploid cell population in most GBM samples.

2.5 Comparison of genomic estimated aneuploidy contents with histologic reports

Histopathologic assessment of tumor purity provides basic information for clinical diagnosis, and is a key criterion in sample selection for research. In TCGA, for example, only GBM with >80% "tumor nuclei" were studied. I found, however, that aneuploid estimates based on SNP data were only moderately correlated with pathologists' report of "percent tumor cells" (Spearman's r = 0.14, P = 0.02, n=275), not correlated with "percent tumor nuclei" (r = 0.076, P = 0.21, n=275), and were lower than AGP by an average of 7% and 18%, respectively (**Figure 2.3**). The difference was not explained by tumors with worse fit in our model, or greater estimation uncertainty. Our inferred AGP is therefore a novel feature extracted from molecular measurements, and can be complementary to the traditionally observed tumor purity.

2.6 Impact of aneuploid content on gene expression patterns

I examined 128 GBM1 samples with both gene expression and CNA data. First, samples of low AGP tend to cluster together in PCA of gene expression data, driving a strong correlation between the first principal component scores (PC1) and AGP (Pearson’s r = 0.62, P =

7.3×10^{-15}, n = 128) (**Figure 2.2C**). PC2 was also correlated with AGP (r = 0.48, P = 1.1×10^{-8}). This pattern suggests that within-tumor heterogeneity is a major driver of gene expression variation, and a factor overlooked in most previous studies. To see if the results for GBM extend to other tumor types, I applied a similar analysis to SNP and expression data for 509 ovarian (OV) tumors from TCGA (The Cancer Genome Atlas Research Network, 2011), and observed a similar pattern (**Figure 2.4)**, with a strong correlation between AGP and PC1 (r = 0.56, P < 2.2×10^{-16}, n = 504). In contrast to AGP, clinically recorded purity values showed little correlation with PC1, r was 0.004 (P = 0.96) for "tumor nuclei", and 0.14 (P = 0.10) for "tumor cells", thus underscoring a key advantage of empirical measures of intra-tumor heterogeneity (Shirahata et al., 2007). Similar to mRNA, expression patterns of 504 microRNAs were also correlated with AGP (r = -0.26, P = 3.0×10^{-3} for PC1; r = -0.56, P = 1.7×10^{-11} for PC2, n=125).

2.7 Combined use of DNA and mRNA patterns in class discovery

The results above raised the question of whether varying levels of euploid-aneuploid mixing could affect the detection of tumor subtypes. To answer this, I performed a joint classification analysis of DNA and mRNA data. In PCA of DNA copy number data, high-AGP samples had high and low PC1s, flanking low-AGP samples **(Figure 2.5A)**, and this was mostly due to a split of Proneural samples (colored purple). Interestingly, PC1 for copy number and PC1 for expression data, when plotted together**,** showed a clear separation of two groups (**Figure**

2.2D), which, due to annotation efforts described below, I will call Non-Proneural and Proneural samples (even though the Proneural group defined here only partially overlaps with the previously defined Proneural group (Verhaak et al., 2010b)). The two groups were not readily separable when either dataset was analyzed by itself. MicroRNA PC2 was highly correlated with PC1 of mRNA data (not shown); thus the joint use of this quantity with copy number PC1 also separated the two groups (**Figure 2.5B**). The Proneural class consisted of 20 high-AGP samples (AGP = 0.86 ± 0.11), of which all but one belonged to the Proneural group defined previously (Verhaak et al., 2010b). Conversely, only 19 out of 38 previously defined Proneural samples (among the 128 analyzed) were Proneural here. Thus, our first revision of GBM classification is that the previously recognized Proneural group splits into two, about half becoming the newly recognized Proneural GBM, another half joining the Non-Proneural class. The Non-Proneural GBMs fell on a continuous distribution that parallels a gradient of AGP (range: 0.23-0.99), and span from the former Mesenchymal samples toward the Classical, Neural, and the rest of the former Proneural samples (**Figure 2.2D**).

I sought to validate these findings in the second batch of GBM (GBM2), using 154 samples having both DNA and mRNA data. AGP estimates were generated as above, showing a similar distribution of AGP in PCA plots of CNA and gene expression data (**Figure 2.6A-B**). Just as in GBM1, combined analysis revealed two well separated classes (**Figure 2.6C)**, with 15 Proneural samples.

2.8 Molecular and clinical features of Proneural GBMs (Proneural/G-CIMP+)

To provide biological annotation of Non-Proneural and Proneural samples, I first note that they carried distinct CNA patterns. Non-Proneural GBMs carried recurrent gains in chromosomes 7, 19, 20, recurrent losses in chromosomes 9p and 10, and a gradient of CNA intensities due to varying AGP (**Figure 2.7A**). Proneural tumors, in contrast, lacked most of the Non-Proneural features described above and had high AGP values. They carried a more diverse set of CNAs, including 11p15.2 deletions (n=12 out of 20), 8q24.21 amplification (n=7), and 10p11.23 amplifications (n=14). Two of the Proneural samples showed co-occurrence of chr1p loss and chr19q loss (bottom of **Figure 2.7A**), each of which was rarely seen in other samples, yet this co-deletion has been reported as a key feature in anaplastic oligodendrogliomas (Cairncross et al., 1998, Ducray et al., 2008). Proneural GBMs had more *IDH1* mutation, a hallmark of secondary GBM (Cooper et al., 2010, Kleihues and Ohgaki, 1999, Nobusawa et al., 2009). They showed higher frequencies of mutations in *TP53*, lower frequencies of mutations in *PTEN*, fewer deletions of *CDKN2A* - these are also signatures of secondary GBM reported previously (Kleihues and Ohgaki, 1999, Ohgaki and Kleihues, 2007). They also showed fewer amplifications and over-expression of *EGFR*, high expression of *PDGFRA*, and lower expression of *FAS* and *MDM2* (**Figure 2.7B and Table 2.2**).

I also compared clinical outcome between the two groups. Compared to Non-Proneural GBM, patients with Proneural GBM were younger at diagnosis (**Figure 2.7C**) and had longer survival time (**Figure 2.7D**). Notably, while the Proneural group defined here has a better

outcome, the other half of the former Proneural group (which I assigned to non-Proneural), is significantly worse than the rest of the Non-Proneural group (P=0.0059). Thus, lumping the two dissimilar types of GBM in the previously defined Proneural class would have missed a clinically relevant distinction.

A recent study of methylation patterns in TCGA samples revealed a subclass of GBM with glioma-CpG island methylator phenotype (G-CIMP+), an epigenetic signature associated with secondary or recurrent GBM and with IDH1 mutations (Noushmehr et al., 2010). Of the 20 Proneural samples I identified, 15 were G-CIMP+ (**Figure 2.7B**); whereas of the 108 Non-Proneural samples none was G-CIMP+, strongly supporting Proneural GBM as a biologically distinct subtype. Indeed, 3-way analysis of CNA, gene expression, and DNA methylation data revealed consistent separation between Proneural and Non-Proneural GBMs (**Figure 2.8**). Proneural samples also match the Proneural GBMs defined in Phillips et al. (Phillips et al., 2006b) (**Table 2.3**). As the term "Proneural" was applied differently in Verhaak et al. and Phillips et al. I renamed the Proneural group as Proneural/G-CIMP+ (or PN/G-CIMP+). PN/G-CIMP+ samples carry signatures resembling those of secondary GBM or low-grade gliomas (Cooper et al., 2010), despite the fact that all but four samples in TCGA have been designated as primary (three of these were PN/G-CIMP+). These results suggest that a fraction (20 of 128 analyzed, ~16%) of the apparently primary GBM cases recruited in TCGA may in fact be latent secondary cases.

2.9 Three subclasses within Non-Proneural GBMs: Molecular and clinical signatures

After Proneural/G-CIMP+ GBMs were recognized, I sought to identify subclasses within the remaining, Non-Proneural GBMs. The reason for removing an already recognized group (i.e., Proneural/G-CIMP+) when studying the fine structure inside another (Non-Proneural) is that the markers distinguishing the two main groups may not be most informative for the within-group analyses, and could confound the latter.

2.9.1 A two-step procedure that relies on GBM1-GBM2 mutual validation

In PCA, Non-Proneural GBMs described a nearly continuous distribution (**Figure 2.2C**), in which the low-AGP samples aggregated to the left, and there were no clearly separated sub-groups in this type of plot. For practical reasons it is often useful to partition seemingly continuously varying samples into discrete classes in order to draw broad biological conclusions, and to aid clinical decision-making. With high-dimension data, however, even samples from a homogeneous distribution can be divided into pre-specified numbers of clusters; but the result can be unstable, and be sensitive to samples included, or the statistical algorithms applied. Self-aggregating algorithms such as hierarchical clustering or k-means clustering will always produce a desired number of clusters; and Consensus Clustering is prone to exaggerate cluster stability (Senbabaoglu et al., manuscript under preparation). In CC, class assignment can be sensitive to outlier samples, chance occurrence of tightly

clustered samples, and the markers used. In addition, as gene-gene correlation is ubiquitously observed, and if groups of highly correlated genes appeared in both the test cohort and validation cohort, it is easy to find that the most discriminating genes in one cohort are "validated" in the second cohort by observing similar clustering patterns.

To address these methodological challenges, I placed major emphases on mutual validation between the GBM1 and GBM2 cohorts rather than first selecting the most informative genes in one and testing them in another. I also focused on the Non-Proneural samples. I ran K-means-based CC on quantile-normalized gene expression data for GBM1, and separately for GBM2, recording the class assignments for K = 2, 3, 4 (K is the number of clusters) for both cohorts. To assess classification concordance between GBM1 and GBM2, I calculated the cross-correlation matrix between every sample in GBM1 and every sample in GBM2, and displayed the resulting matrix where samples were grouped by class assignments independently obtained for the two cohorts (**Figure 2.9A-B**). If samples of a given class in GBM1 showed high correlation coefficients (r) with those of a particular class in GBM2, and showed low r values with other GBM2 classes, the class discovery was considered mutually validated. Conversely, if the classes did not show a one-to-one correspondence between the two independent cohorts, I considered the class definition poorly replicated. **Figure 2.9A** showed the GBM1-GBM2 cross-correlation matrix for K=2, where the two classes defined in GBM1 could be matched, one-to-one, to the two classes independently defined in GBM2. In comparison, K=3 or 4 yielded substantially worse matching (**Figure 2.10A-B**).

At K=2, one of the two classes for GBM1 contained all the 37 samples in the Mesenchymal group defined previously (**Figure 2.9A**). I therefore named it the Mesenchymal (MES) group

even though it now also contained 4 former Neural/Proneural samples and 11 former Classical samples. The other class showed hints of finer structure in **Figure 2.9A**; and this was explored by repeating the analysis described above within this class. This led to a further split into 2 subclasses (n=27 and 29, **Figure 2.9B**), with K=2 being better than k=3 or 4 (**Figure 2.10C-D**). One of the subclasses was dominated by the previously defined Classical samples, and was thus named the Classical group even though it also contained 5 Neural and 1 Proneural samples. The other subclass, with a mixture of Non-Proneural-Proneural and Neural tumors, was named Proliferative for its similarity with the Proliferative samples identified by Phillips et al. (Phillips et al., 2006b). Attempts to identify further subtypes within the Proliferative group were not supported by mutual validation between GBM1 and GBM2 (not shown). This led us to conclude that the G-CIMP-minus (G-CIMP-) subset of previously defined Proneural samples did not form a distinct group. In other words, there wasn't a second, self-contained Proneural group in the current GBM dataset, although it is possible that a larger sample size in future studies could have the power to reveal finer splits. In all, I identified three subclasses for Non-Proneural GBM through a two-stage, stepwise clustering procedure, with optimal K=2 at both stages, and supported by concordance between GBM1 and GBM2. The resulting assignments were different (by 12% of samples) from those assigned by a one-stage, K=3 approach. I consider the two-stage approach more appropriate because the finer division in the second stage is not affected by the more divergent profiles of the two main classes identified in the first stage. The three newly identified Non-Proneural GBM classes are visually coherent on the gene expression PCA plot, for both GBM1 (**Figure 2.9C**) and GBM2 (**Figure 2.9D**).

2.9.2 Comparison with previous studies

Phillips et al. (Phillips et al., 2006b) proposed a three-class system for GBM: Proneural, Proliferative and Mesenchymal. Verhaak et al. (Verhaak et al., 2010b) reported four classes for TCGA samples: Proneural, Neural, Classical, and Mesenchymal; and in this work I described a revised four-class system for the same dataset as in Verhaak et al. Based on molecular signatures and comparisons with Phillips et al and Verhaak et al's work discussed below, I name the three revised Non-Proneural GBM classes: Classical, Proliferative and Mesenchymal. In order to summarize how these systems have evolved (i.e., how different classes correspond to each other), I first reanalyzed the Phillips' data, which were made publicly available and a subset of 56 samples were subsequently processed to combine two technical platforms (Verhaak et al., 2010b). Among the 56 samples I first observed that the Proneural samples in Phillips' study showed high similarities to our PN/G-CIMP+ GBMs in terms of patient age, survival time, and patterns of CNAs (not shown). For the remaining 46 samples, which were designated Non-Proneural GBMs here, I followed the procedure of Phillips et al. to select 584 genes most highly correlated with patients' survival time (out of 1,740 most variable genes) and performed k-means clustering, using cross-correlation with TCGA's GBM1 to find the optimal number of classes. Again, K=2 yielded the best match for both steps in a two-step procedure (**Figure 2.11**), leading to the recognition of 19 Mesenchymal, 14 Proliferative and 13 Classical samples. This new three-way classification

of 46 Non-Proneural samples showed better cohesion on the PCA plot (**Figure 2.12**) than the original classification, and this could be explained by noting that the latter was based on a different set of (and much fewer) genes.

By tracking the class reassignments between the two datasets and between the original classification and our revised classification (**Figure 2.13A-B**), I documented the commonalities and differences among different classification systems (**Figure 2.13C, Table 2.3**). Of the 108 samples, 70 (65%) had one-to-one mapping to the previous NL, CL, and MES classes (**Figure 2.13**); thus 35% of GBM1 samples received revised assignments. I similarly analyzed the 46 Non-Proneural samples in Phillips et al. **(Figure 2.11-12)**, and found that the former Proliferative group was split into the new Proliferative and Classical groups, and 11 (24%) were reassigned into or out of the MES group. The MES class was reproducibly identified in both datasets and in both the original and the revised schemes. The original Proneural group for TCGA was split into (1) the PN/G-CIMP+ group, which is equivalent to Phillips' Proneural group, and (2) Non-Proneural-Proneural (N-P-P), which was merged with the original Neural samples to form the revised Proliferative group, which closely resembles Phillips' Proliferative group. However, some of Phillips' Proliferative samples split and formed the revised Classical group, which closely resembles the original Classical group for TCGA samples. In sum, a major revision of the Verhaak et al. classification is in recognizing that the Proneural group contains two distinct subgroups, one of which, PN/G-CIMP+, is well separated from the other three classes by CNA patterns, IDH1 mutations, patient age, and outcome. For the Philipps' dataset, a major revision is in separating the original Proliferative group into the revised Proliferative and Classical groups.

The primary reason that the Proneural/G-CIMP+ class was previously mis-grouped with some Neural samples is that their gene expression signatures, when viewed without other genomic data, were not sufficiently distinctive, because Proneural samples share a cell type-specific signature with the Neural samples (renamed as the Proliferative samples in our system). It was only by integrating the CNA data (this work) or by using the methylation data (Noushmehr et al., 2010) that the Proneural/G-CIMP+ group became evident. The Phillips' study did not miss this group because the authors selected genes strongly correlated with survival time rather than those showing the largest variation. Since patients in the Proneural/G-CIMP+ group survived longer, genes that were most informative for recognizing this group were used in that study.

2.9.3 Clinical relevance of revised Non-Proneural GBM classes

Since any new method could lead to a different classification, I pursued an important question: are the biological features of the new classes more robust than in the old system? Many marker genes highlighted in previous studies were consistently observed (**Table 2.4**). In CNA patterns (**Figure 2.14A**), while Non-Proneural samples shared the chr7 gains and chr10 losses, Proliferative samples carried additional deletions in chr14 and chr15 rarely seen in Classical samples (Student t test for chromosome-wide averaged copy number: $P=2.6\times10^{-3}$ and 3.1×10^{-4}, for chr14 and chr15, respectively), whereas Classical samples carried more amplifications in chr19 ($P=1.1\times10^{-6}$) and chr20 ($P=3.6\times10^{-6}$) than in Proliferative samples.

Interestingly, many MES samples carried both the chr14-15 deletions and the chr19-20 gains, although with varying intensities due to lower aneuploid content, and with significantly more chr13 deletions compared with non-MES tumors ($P=9.6\times10^{-3}$). For Proliferative and MES samples, chr14q and 15 deletions tended to be mutually exclusive (mean Pearson's r = -0.23 for Prolif and -0.26 for MES); whereas for Classical samples, chr19 gains tended to co-occur with chr20 gains (mean Pearson's r = 0.46). These results showed that in addition to the CNA differences between Non-Proneural and PN/G-CIMP+ (**Figure 2.7A**), the three Non-Proneural classes carried different patterns of genomic aberration, possibly reflecting their differences in cell lineage, transcriptome patterns, and patient outcome.

The three Non-Proneural classes also showed significant differences in survival time in a three-way comparison in GBM1 (**Figure 2.14B,** log-rank test P=0.011). This is in contrast to the previous class assignments (Verhaak et al., 2010b), for which the three-way comparison was not significant (**Figure 2.14C**). For individual pairs of classes, five out of six pairwise comparisons were significant in the revised system, while only one of six was significant in the previous system (**Figure 2.15**). The revised classes for Phillips' dataset also had significant survival differences in the three-way comparison (P = 0.033, log-rank test) and in the four-way comparison that included the PN/G-CIMP+ group (P = 0.014).

To directly compare the relative hazard across the four GBM subtypes and incorporate relevant patient characteristics, I performed a Cox proportional hazard regression analysis using our four-class assignments as explanatory covariates, and including patient age and the Karnofsky Performance Status (KPS) scores. First, for the entire set of 128 GBM1 samples, with the PN/G-CIMP+ subtype used as the reference category, the three non-Proneural

subtypes had higher hazard ratios in the revised system (**Figure 2.16A**) than in the previous system (**Figure 2.16B**). Second, when I focused only on the three non-Proneural subtypes, using Classical as the reference, the 108 samples in the revised system (**Figure 2.16C**) showed higher hazard ratios than the 98 samples in the previous system (**Figure 2.16D**). To compare concordance between tumor classification and patient outcomes, I computed the C-statistics (Harrell et al., 1996) for the 128 GBM1 dataset using the Cox regression model with age, KPS and subtypes as covariates. Revised classification had a concordance score of 0.668, higher than using age and KPS alone (0.643) by 2.5%, whereas the previous system had a concordance of 0.651, higher than using age and KPS alone (0.643) by only 0.8%, indicating that the revised system had improved predictive power for patient outcome.

2.9.4 Validation of survival time differences in an independent cohort

The Non-Proneural classes described above were defined by mutual validation of GBM1 and GBM2, thus having used information from both cohorts. To validate the survival time differences in a new, independent dataset, I analyzed a third batch of 144 TCGA samples (GBM3). As before, I identified 26 PN/G-CIMP+ samples using expression data and CNA data. Survival time differences were indeed validated in GBM3, with five out of six pairwise comparisons showing significant differences (**Table 2.5A**). To compare with the previous system, I used the 840 markers suggested by Verhaak et al. to classify the GBM3 samples and found that only one of six pairwise comparisons was significant (**Table 2.5B**).

2.10 Inference of cell type composition of GBM classes

I attempted to deduce the possible cell type composition of the four GBM classes to shed light on the cellular origins of this heterogeneous cancer. To do so, I compared GBM expression data with a reference dataset, GSE9566 (Cahoy et al., 2008), for 38 samples that represent four main cell types in the central nervous system: acutely isolated astrocytes, neurons, oligodendrocytes, and cultured astroglia. The 38 samples formed four well-separated clusters, in agreement with their known identity (**Figure 2.16**). Cross-correlations of Non-Proneural GBM samples with the 38 reference samples, when grouped by class (for GBM) and cell type (for reference samples), showed recognizable mapping of GBM classes to known neural cell types, for GBM1 (N=128), GBM2 (N=154), and Phillips' dataset (N=56) (**Figure 2.17A-C**). Both PN/G-CIMP+ and Proliferative samples showed high correlations with neurons and oligodendrocytes, suggesting that they both resemble oligodendrogliomas. The Classical samples were similar to the astrocytes, suggesting that they may be related to astrocytomas. Lastly, the Mesenchymal samples showed high similarities with the cultured astroglia samples, which had an "immature or reactive phenotype" (Cahoy et al., 2008), consistent with the MES signatures of angiogenesis and inflammatory infiltration (Phillips et al., 2006b, Verhaak et al., 2010b, Murat et al., 2009). The observed resemblance to known cell types was generally consistent with what was reported previously (Verhaak et al., 2010b), but with important differences. First, the former

Neural group did not show clear mapping to any cell type. Second, the mapping to reference cell types is much stronger with the new system: the difference (D) of the mean correlation coefficients between the mapped diagonal blocks and the off-diagonal blocks of the correlation matrix (**Figure 2.17A-C**) was 0.562 for the new classes, much higher than in the previously reported classes (D = 0.247) even when I counted the best mapped blocks for the latter.

As most of the low-AGP samples fell in the Mesenchymal group, I attempted to clarify the cell lineage of the aneuploid and euploid populations. If the aneuploid cells were derived from one of the reference cell types, there should be a positive correlation between (1) the correlation between samples of that particular cell type and individual MES tumors and (2) the MES tumors' AGP values, which measure how much aneuploid cells they contain. I calculated the correlation coefficients r, for each of the 38 reference samples, between its correlation coefficients with the MES samples and the AGP values of the MES samples, and found consistent and positive r values for Cultured Astrocytes (**Figure 2.17D**), suggesting that the aneuploid cells in MES share gene expression features, and possible common lineage, with reactive astrocytes (Cahoy et al., 2008).

As no other cell type in the reference set showed negative correlations, the identity of the euploid cells in MES remained unexplained. MES tumors carry angiogenic and inflammatory signatures, and some microglia markers are highly expressed in MES samples (Verhaak et al., 2010b). I therefore hypothesize that the euploid fraction may be related to microglia/macrophage infiltration. To test this hypothesis, I searched public databases for gene expression data for microglia samples, and found data for tumor-infiltrating

microglia/macrophage isolated from freshly excised brain tumors ("TI. microglia", in GSE25289) (Mora et al., 2010) and for microglia fraction from postoperative GBM tissue ("G. microglia", in GSE16119) (Murat et al., 2009). The correlation of these cells with MES tumors showed negative correlations with AGP (**Figure 2.4D**), suggesting that expression signatures of MES euploid cells are similar to microglia/macrophage. Moreover, two microglia/macrophage-specific transcripts, integrin alpha M (*ITGAM*) (Guillemin and Brew, 2004) and allograft inflammatory factor-1 (*AIF1*) (Schwab et al., 2001), were negatively correlated with AGP ($r = -0.58$, $P = 6.3\times10^{-6}$ for *ITGAM*; $r = -0.53$, $P = 5.3\times10^{-5}$ for *AIF1*), further supporting microglia/macrophage as the probable source of euploid population in MES.

2.11 Hierarchical classification of GBM

The new understanding of GBM genomic landscape led to our proposal of a cohesive stepwise classification procedure (**Figure 2.18**). First, Proneural/G-CIMP+ GBMs can be identified with joint analyses of copy number and mRNA profiles, along with clinical data such as patient age. Even if a case was recorded as primary GBM due to the apparent lack of antecedent tumors, it could be recognized as Proneural/G-CIMP+ by features such as younger age, *IDH1* mutations, lack of *PTEN* mutations, hyper-methylation patterns, and lack of chr7 gains and chr10 losses. Among the remaining, Non-Proneural samples, MES samples can be separated from the Classical and Proliferative samples by lower AGP values, necrosis

signatures, higher expression of *FAS* and *CHI3L1*, etc. These tumors experienced more infiltration of non-cancerous cells, containing aneuploid reactive astrocyte-like cells intermingled with cells such as microglia/macrophage that lack CNAs. Lastly, Classical and Proliferative samples can be distinguished by gene expression patterns that resemble different neural cell types. Known markers highlighted by previous studies **(Table S4)**, such as *PCNA* and *TOPA2A* overexpression in Proliferative samples, can also be incorporated in this step.

2.12 Summary

The practice in this work covered inter-tumor level and sample level heterogeneity. I have developed an algorithm to estimate the euploid cell mixing ratios in surgically removed bulk tumor tissues. My algorithm falls in the lineage of pattern recognition discussed in **Chapter 1**, **4.2**, which was first introduced by Popova et al (ref). However, AGP inference algorithm differentiates from the original GAP method in several ways: first, it used folded BAF as x-axis, which allowed me to introduce contraction tracks for each possible CNA configuration. This feature will later be used to develop a more capable tool, segmental AGP inference algorithm, which estimates

estimation. Second, AGP inference algorithm does not rely on external input of tumor DNA content to genotype the somatic CNAs. It is able to estimate tumor and sample ploidy based on the distribution of observed data points on the BAF-LRR plot.

Discoveries of GBM subtypes have so far relied on single data types. The work reported here combined DNA genotyping data and gene expression data, and revealed a novel GBM subtype (Proneural/G-CIMP+) that carried distinct molecular, clinical, and demographic features. While this subtype was described separately in a study of methylation data (Noushmehr et al., 2010), our approach reached the conclusions from two other, independent data types, and suggests that such a combined approach will be useful in genomic analysis of other cancers.

Table 2.1: Genomic features and QC measures for GBM1 and GBM2 (n=284)

Tumor ID	AGP	std[1]	PC[2]	PoP[3]	2n.4n.6n[4]	pl.tumor[5]	pl.overall[6]	%amp[7]	%del[8]	%del.loh[9]	%cn.loh[10]	2.5%[11]	median[12]	97.5%[13]	%TumorCell[14]	%TumorNuclei
TCGA-02-0001	0.68	0.0811	0.837	0.867	4	4.21	3.5	0.0954	0.742	0	0	0.51475	0.55	0.68	90	95
TCGA-02-0003	0.92	0.244	0.159	0.853	2	1.94	1.95	0.0522	0.106	0.0838	0	0.13	0.86	0.95	100	100
TCGA-02-0006	0.59	0.0239	0.272	0.844	2	1.95	1.97	0.101	0.17	0.131	0	0.55	0.6	0.64	95	100
TCGA-02-0007	0.96	0	0.251	0.928	2	1.86	1.87	0.0361	0.215	0.17	0.0281	0.05	0.05	0.05	90	100
TCGA-02-0009	0.93	0.0302	0.18	0.871	2	1.96	1.96	0.0735	0.107	0.0834	0	0.86	0.925	0.96	100	100
TCGA-02-0010	0.86	0.106	0.439	0.647	4	4.05	3.77	0.0878	0.351	0	0.0457	0.56475	0.78	0.95	100	100
TCGA-02-0011	0.95	0.238	0.344	0.981	4	3.73	3.65	0.0513	0.293	0	0	0.27	0.8	0.9	95	100
TCGA-02-0014	0.98	0.014	0.259	0.888	2	1.95	1.96	0.0913	0.167	0.112	0.0264	0.95	0.98	0.99	100	100
TCGA-02-0015	0.92	0.0261	0.519	0.486	4	4.18	4	0.227	0.291	0	0	0.7795	0.84	0.88	80	100
TCGA-02-0016	0.88	0.0104	0.672	0.878	4	4.2	3.94	0.317	0.354	0	0	0.79	0.8	0.8	50	80
TCGA-02-0021	0.9	0.00902	0.202	0.959	2	2.01	2.01	0.11	0.0922	0.0844	0	0.87	0.89	0.9	90	100
TCGA-02-0023	0.75	0.139	0.935	0.907	4	4.75	4.06	0.572	0.363	0	0.0562	0.42475	0.68	0.76525	100	100
TCGA-02-0024	0.92	0.234	0.506	0.995	4	3.94	3.79	0.123	0.383	0	0	0.24475	0.475	0.87	80	95
TCGA-02-0025	0.36	0.00716	0.968	0.992	4	8.24	4.25	0.701	0.267	0	0.0349	0.34	0.36	0.36525	80	100
TCGA-02-0026	0.9	0.0196	0.314	0.755	6	5.94	5.54	0.06	0.254	0	0	0.75	0.78	0.82	80	100
TCGA-02-0027	0.75	0.139	0.174	0.992	6	7.14	5.86	0.0557	0.118	0	0	0.25	0.49	0.76	80	100
TCGA-02-0028	0.97	0.246	0.25	0.999	4	4.09	4.02	0.146	0.104	0	0	0.36275	0.795	0.99	80	100
TCGA-02-0033	0.41	0.147	0.177	0.994	2	1.94	1.98	0.0553	0.122	0.121	0	0.07	0.4	0.43	80	100
TCGA-02-0034	0.48	0.0507	0.258	0.845	2	1.87	1.94	0.0656	0.193	0.161	0	0.43475	0.47	0.51	90	100
TCGA-02-0037	0.87	0.0763	0.233	0.985	2	2.06	2.05	0.142	0.0905	0.0654	0.0219	0.79475	0.87	0.89	80	100
TCGA-02-0038	0.82	0	0.199	0.986	2	1.97	1.98	0.0897	0.11	0.108	0	0.05	0.05	0.05	100	100
TCGA-02-0039	0.68	0.00732	0.331	0.873	6	7.39	5.67	0.106	0.225	0	0	0.42	0.43	0.44525	80	100
TCGA-02-0043	0.99	0.134	0.2	0.565	4	4.01	3.99	0.101	0.0995	0	0	0.4495	0.96	0.99	80	100

TCGA-02-0051	0.79	0.00768	0.84	0.934	4	4.81	4.22	0.403	0.437	0	0	0.66	0.66	0.69	80	100
TCGA-02-0052	0.71	0.0948	0.19	0.526	2	1.9	1.93	0.0177	0.172	0.0987	0	0.38375	0.71	0.7405	85	100
TCGA-02-0054	0.49	0.0953	0.573	0.884	4	6.56	4.24	0.363	0.21	0	0	0.19325	0.295	0.5505	95	95
TCGA-02-0055	0.69	0.0744	0.667	1	8	9.54	7.2	0.0635	0.603	0	0	0.1	0.35	0.48	80	95
TCGA-02-0057	0.39	0.131	0.517	0.968	2	1.96	1.98	0.195	0.322	0.124	0	0.08	0.38	0.39	80	100
TCGA-02-0058	0.85	0.18	0.0659	0.908	4	4.24	3.9	0	0.0659	0	0	0.37	0.74	0.83	80	95
TCGA-02-0059	0.68	0.0804	0.369	0.948	6	7.59	5.8	0.11	0.259	0	0	0.41	0.42	0.68	80	100
TCGA-02-0060	0.73	0.272	0.189	0.399	2	1.96	1.97	0.0109	0.178	0.0649	0	0.13	0.23	0.79	60	90
TCGA-02-0064	0.67	0.00792	0.196	0.978	2	1.95	1.96	0.0752	0.121	0.117	0	0.65	0.66	0.68	95	100
TCGA-02-0068	0.76	0.039	0.351	0.911	4	4.6	3.97	0.162	0.189	0	0	0.63	0.76	0.77	80	100
TCGA-02-0069	0.92	0.0497	0.923	0.839	4	3.35	3.24	0.138	0.785	0	0	0.78	0.845	0.93	75	95
TCGA-02-0070	0.73	0.0737	0.877	0.984	4	4.88	4.1	0.727	0.151	0	0	0.57	0.73	0.73	95	100
TCGA-02-0071	0.7	0.0975	0.18	0.862	2	2	2	0.0927	0.0874	0.0563	0	0.28275	0.73	0.74	85	100
TCGA-02-0074	0.68	0.172	0.531	0.945	6	7.11	5.48	0.0584	0.473	0	0.003	0.31	0.66	0.8	80	100
TCGA-02-0075	0.67	0.224	0.564	0.94	6	7.23	5.51	0.0599	0.504	0	0	0.07	0.42	0.837	90	100
TCGA-02-0079	0.49	0.00591	0.37	0.964	4	5.8	3.86	0.129	0.241	0	0	0.32	0.33	0.34	50	90
TCGA-02-0080	0.76	0.0877	0.143	0.846	4	4.78	4.11	0.0894	0.0533	0	0	0.54	0.735	0.82525	80	100
TCGA-02-0083	0.92	0.00332	0.15	0.907	2	1.93	1.94	0.0439	0.106	0.0956	0	0.91	0.92	0.93	80	100
TCGA-02-0084	0.78	0.132	0.388	0.997	6	6.83	5.77	0.0874	0.301	0	0	0.23475	0.55	0.8605	95	100
TCGA-02-0085	0.47	0	0.13	0.974	2	1.98	1.99	0.056	0.0745	0.0711	0	0.05	0.05	0.05	75	90
TCGA-02-0086	0.71	0.00682	0.91	1	4	3.71	3.22	0.0551	0.855	0	0	0.53	0.54	0.55	80	100
TCGA-02-0089	0.71	0.00595	0.182	0.919	2	1.98	1.99	0.0871	0.0952	0.0839	0	0.69	0.7	0.71	90	100
TCGA-02-0099	0.75	0.00498	0.329	0.987	2	1.86	1.89	0.0712	0.258	0.214	0.0423	0.74	0.75	0.75	90	100
TCGA-02-0102	0.93	0.172	0.139	0.966	2	2.02	2.01	0.0737	0.065	0.0619	0	0.36	0.84	0.94	80	95
TCGA-02-0104	0.95	0.0176	0.35	0.166	4	3.99	3.89	0.151	0.199	0	0	0.89	0.91	0.95	95	100
TCGA-02-0107	0.23	0	0.19	1	2	1.89	1.97	0.0454	0.145	0.145	0.0244	0.05	0.05	0.05	80	90
TCGA-02-0113	0.93	0.0568	0.613	0.999	8	6.93	6.58	0.00051	0.612	0	0	0.36	0.37	0.38	90	95
TCGA-02-0114	0.98	0.127	0.24	0.999	4	3.91	3.87	0.106	0.134	0	0.0149	0.50325	0.96	0.99	90	100

TCGA-02-0115	0.52	0.095	0.364	0.994	4	5.64	3.89	0.106	0.258	0	0	0.16425	0.36	0.48	100	100
TCGA-02-0116	0.95	0.0875	0.217	0.934	2	1.98	1.98	0.103	0.114	0.101	0	0.84	0.95	0.98	80	100
TCGA-06-0122	0.76	0.0122	0.261	0.995	2	1.95	1.97	0.0927	0.168	0.14	0.0276	0.73	0.76	0.78	80	100
TCGA-06-0124	0.59	0.00522	0.2	0.989	2	1.93	1.96	0.0674	0.133	0.131	0	0.58475	0.59	0.6	100	100
TCGA-06-0125	0.93	0.0196	0.18	0.956	2	2.02	2.02	0.101	0.0784	0.0574	0.016	0.85475	0.93	0.94	95	100
TCGA-06-0126	0.82	0.18	0.191	0.927	2	2.01	2.01	0.102	0.089	0.0762	0	0.15	0.83	0.86	95	100
TCGA-06-0127	0.62	0.00456	0.13	0.956	2	2.01	2.01	0.0624	0.0674	0.0494	0.0129	0.62	0.62	0.63	80	100
TCGA-06-0128	0.82	0.248	0.244	0.64	6	6.56	5.74	0.038	0.206	0	0	0.01	0.275	0.86	95	100
TCGA-06-0129	0.93	0.0382	0.233	0.813	4	3.96	3.82	0.0201	0.213	0	0	0.76	0.81	0.9	100	100
TCGA-06-0130	0.36	0.0729	0.554	0.998	4	7.13	3.85	0.119	0.435	0	0.0103	0.18	0.33	0.37	95	95
TCGA-06-0132	0.35	0.0818	0.416	0.999	4	7.7	4	0.229	0.187	0	0	0.07	0.18	0.33525	90	100
TCGA-06-0137	0.93	0.105	0.213	0.792	2	2.02	2.02	0.133	0.0795	0.0643	0.00426	0.82425	0.94	0.96	85	100
TCGA-06-0138	0.66	0.0474	0.207	0.726	2	1.98	1.99	0.0926	0.114	0.0896	0	0.56	0.64	0.67	85	100
TCGA-06-0139	0.09	NA	0.00201	1	2	2.01	2	0	0.00201	0.00201	0.00201	NA	NA	NA	50	95
TCGA-06-0143	0.73	0.157	0.177	0.786	2	1.98	1.98	0.0728	0.104	0.0835	0	0.16	0.71	0.76525	75	95
TCGA-06-0145	0.79	0.00664	0.173	0.975	2	1.99	1.99	0.0842	0.0887	0.0883	0	0.78	0.8	0.81	80	100
TCGA-06-0147	0.24	0.0481	0.258	0.997	4	10.5	4.04	0.18	0.0775	0	0	0.06	0.14	0.24	60	100
TCGA-06-0148	0.93	0.0238	0.222	0.936	2	2	2	0.0982	0.124	0.0586	0.0523	0.86	0.93	0.95	90	100
TCGA-06-0150	0.62	0.0505	0.606	0.978	4	5.28	4.03	0.224	0.382	0	0	0.34	0.44	0.47	0	50
TCGA-06-0152	0.8	0.00729	0.582	0.997	4	4.22	3.78	0.161	0.421	0	0.000265	0.66	0.67	0.68	68	95
TCGA-06-0154	0.74	0.00623	0.226	0.804	2	1.97	1.98	0.12	0.105	0.0899	0.00949	0.73	0.74	0.75	60	90
TCGA-06-0155	0.81	0.0962	0.144	0.966	2	2	2	0.0772	0.0666	0.0658	0.000671	0.5405	0.82	0.83	80	95
TCGA-06-0156	0.51	0.0641	0.785	0.629	4	5.73	3.9	0.332	0.452	0	0.0607	0.17475	0.34	0.48525	0	0
TCGA-06-0157	0.83	0.198	0.465	0.656	4	4.45	4.03	0.199	0.266	0	0	0.33	0.69	0.85	80	95
TCGA-06-0158	0.8	0	0.224	0.996	4	4.37	3.89	0.0911	0.133	0	0	0.01	0.01	0.01	80	95
TCGA-06-0160	0.05	NA	0	NA	2	2	2	0	0	0	0	NA	NA	NA	0	0
TCGA-06-0165	0.05	NA	0	NA	2	2.01	2	0	0	0	0	NA	NA	NA	0	0
TCGA-06-0166	0.53	0.134	0.503	0.989	6	8.68	5.54	0.0469	0.456	0	0	0.26	0.53	0.54525	20	90

TCGA-06-0167	0.05	NA	0	NA	2	2	2	0	0	0	0	NA	NA	NA	0	0
TCGA-06-0168	0.61	0.0769	0.322	0.937	4	5.22	3.96	0.15	0.172	0	0.0875	0.41	0.47	0.64525	55	85
TCGA-06-0169	0.53	0.143	0.243	0.988	4	5.77	4	0.0814	0.162	0	0	0.14	0.36	0.54	80	100
TCGA-06-0171	0.64	0.124	0.373	0.965	4	4.9	3.85	0.11	0.264	0	0	0.23	0.46	0.66525	50	90
TCGA-06-0173	0.87	0.18	0.214	0.871	4	4.33	4.03	0.126	0.0882	0	0	0.33	0.75	0.85	75	100
TCGA-06-0174	0.88	0.0399	0.144	0.952	2	2	2	0.0774	0.067	0	0	0.75475	0.88	0.89	90	100
TCGA-06-0176	0.59	0.00474	0.988	1	8	9.23	6.27	0	0.988	0	0	0.26	0.27	0.27	55	95
TCGA-06-0184	0.54	0.0103	0.375	0.998	2	1.8	1.89	0.0432	0.332	0.152	0.0583	0.52475	0.54	0.56	75	95
TCGA-06-0185	0.89	0.178	0.254	0.973	6	6.43	5.95	0.13	0.123	0	0	0.2295	0.71	0.99	95	100
TCGA-06-0187	0.87	0.0173	0.197	0.984	2	1.96	1.97	0.0798	0.117	0.116	0	0.8195	0.87	0.89	90	100
TCGA-06-0188	0.83	0.146	0.408	0.993	6	6.68	5.88	0.183	0.225	0	0	0.25475	0.645	0.7505	80	95
TCGA-06-0189	0.17	0	0.0012	1	2	1.91	1.99	0.0309	0.0951	0.126	0.111	0.05	0.05	0.05	75	100
TCGA-06-0192	0.55	0.178	0.276	0.987	4	5.65	4.01	0.101	0.175	0	0	0.14	0.34	0.9805	75	95
TCGA-06-0195	0.82	0.0799	0.7	0.997	4	3.68	3.38	0.057	0.643	0.0189	0.0914	0.64475	0.83	0.85	85	85
TCGA-06-0197	0.51	0.0911	0.239	0.889	2	1.91	1.95	0.0551	0.183	0.0498	0	0.15	0.525	0.58	95	95
TCGA-06-0201	0.45	0.07	0.162	0.99	2	1.97	1.99	0.0724	0.0898	0.0882	0	0.213	0.445	0.5105	60	95
TCGA-06-0206	0.6	0.178	0.206	0.853	4	5.36	4.02	0.0985	0.108	0	0	0.32	0.43	0.99	88	98
TCGA-06-0208	0.72	0.21	0.15	0.965	2	1.95	1.96	0.0534	0.0964	0.0919	0	0.11	0.73	0.74	80	100
TCGA-06-0208	0.72	0.182	0.147	0.965	2	1.96	1.97	0.0521	0.0944	0.09	0	0.1295	0.72	0.75525	NA	NA
TCGA-06-0209	0.58	0.119	0.437	0.926	4	5.47	4.01	0.19	0.247	0	0	0.15475	0.44	0.58	90	100
TCGA-06-0210	0.66	0.0885	0.171	0.988	2	1.95	1.97	0.06	0.111	0.11	0.000285	0.34475	0.66	0.69	85	95
TCGA-06-0210	0.44	0.0639	0.167	0.991	2	1.95	1.98	0.0584	0.109	0.107	0.000281	0.3585	0.44	0.49	NA	NA
TCGA-06-0211	0.93	0.15	0.247	0.701	6	6.39	6.08	0.169	0.078	0	0	0.41	0.66	0.9105	50	95
TCGA-06-0211	0.96	0.0999	0.43	0.537	4	4.08	4	0.186	0.244	0	0	0.7995	0.91	0.96525	NA	NA
TCGA-06-0213	0.71	0.143	0.149	0.545	4	4.82	4	0.0583	0.0903	0	0	0.32	0.56	0.9605	50	95
TCGA-06-0214	0.68	0.138	0.261	0.979	6	7.5	5.74	0.0576	0.203	0	0	0.41	0.67	0.68	80	95
TCGA-06-0216	0.93	0.0613	0.627	0.6	4	3.46	3.36	0.0581	0.569	0.000452	0.157	0.77	0.92	0.95	80	90
TCGA-06-0219	0.64	0.00935	0.382	0.985	2	1.66	1.78	0.0187	0.363	0.362	0	0.62	0.64	0.65525	95	100

TCGA-06-0221	0.75	0.0967	0.219	0.916	4	4.57	3.93	0.043	0.176	0	0.0262	0.43475	0.54	0.75525	50	80
TCGA-06-0237	0.84	0.046	0.288	0.96	4	4.25	3.89	0.0846	0.203	0	0.0459	0.69475	0.8	0.88	95	100
TCGA-06-0238	0.76	0.0264	0.209	0.998	4	4.56	3.94	0.0593	0.15	0	0.0612	0.6775	0.75	0.76	95	100
TCGA-06-0240	0.85	NA	0.000723	0.591	2	2	2	0.000723	0	0	0	NA	NA	NA	0	10
TCGA-06-0241	0.92	0.00533	0.186	0.944	2	1.93	1.94	0.0596	0.127	0.111	0.000219	0.91	0.92	0.92	90	100
TCGA-06-0644	0.61	0.0324	0.318	0.791	4	5.22	3.97	0.0861	0.232	0	0.0942	0.42	0.45	0.5225	70	95
TCGA-06-0645	0.65	0.096	0.244	0.867	4	5.28	4.13	0.193	0.0508	0	0	0.429	0.65	0.67525	80	100
TCGA-06-0646	0.88	0.0521	0.211	0.992	6	6.32	5.8	0.0561	0.155	0	0	0.67	0.71	0.84525	60	90
TCGA-06-0648	0.93	0.0413	0.223	0.936	4	4.1	3.96	0.0917	0.131	0	0	0.79	0.88	0.91	90	100
TCGA-06-0649	0.65	0.148	0.559	0.903	6	7.55	5.61	0.0986	0.461	0	0	0.28	0.64	0.67	70	90
TCGA-06-0686	0.73	0.0871	0.365	0.68	4	4.82	4.05	0.128	0.237	0	0.00116	0.298	0.545	0.64	75	95
TCGA-06-0743	0.8	0	0.274	0.602	2	2.01	2.01	0.126	0.148	0.000426	0	0.05	0.05	0.05	95	100
TCGA-06-0744	0.9	0.0296	0.221	0.759	2	1.97	1.97	0.0964	0.124	0.0851	0	0.81475	0.9	0.93	95	100
TCGA-06-0745	0.82	0.206	0.294	0.852	4	4.26	3.85	0.0905	0.204	0	0.00475	0.15425	0.69	0.76	95	100
TCGA-06-0747	0.92	0	0.285	0.998	2	1.96	1.96	0.0861	0.199	0.092	0.107	0.05	0.05	0.05	100	100
TCGA-06-0749	0.57	0.0906	0.219	0.999	4	5.36	3.92	0.0853	0.133	0	0	0.35	0.555	0.58	60	80
TCGA-06-0750	0.75	0.0116	0.205	0.967	2	1.9	1.93	0.0564	0.148	0.143	0	0.72	0.75	0.76	80	90
TCGA-06-0875	0.92	0.0745	0.272	0.897	4	3.89	3.74	0.0295	0.243	0	0	0.7995	0.89	0.99	100	95
TCGA-06-0876	0.85	0.158	0.199	0.891	6	6.56	5.88	0.0508	0.148	0	0	0.10475	0.65	0.68525	100	95
TCGA-06-0877	0.8	0.315	0.268	0.99	2	2.13	2.11	0.205	0.063	0.0588	0	0.13	0.8	0.81	95	95
TCGA-06-0878	0.83	0.00931	0.286	0.813	2	1.9	1.92	0.096	0.19	0.165	0	0.81	0.83	0.84	95	95
TCGA-06-0881	0.83	NA	0.000643	1	2	2	2	0.000643	0	0	0	NA	NA	NA	70	95
TCGA-06-1084	0.55	0.102	0.327	0.952	4	5.56	3.96	0.147	0.18	0	0	0.15	0.34	0.56	80	90
TCGA-06-1086	0.97	0.0835	0.735	0.995	4	3.32	3.28	0.0324	0.703	0	0.0643	0.65475	0.93	0.97	75	85
TCGA-06-1087	0.99	0.0194	0.64	0.696	4	3.43	3.41	0.114	0.526	0	0.0635	0.92	0.97	0.99	85	90
TCGA-06-1800	0.83	0.251	0.281	1	6	6.51	5.74	0.0551	0.226	0	0	0.21	0.62	0.99	65	80
TCGA-06-1801	0.83	0.0765	0.609	0.99	4	4.3	3.91	0.352	0.257	0	0	0.58825	0.83	0.84525	80	85
TCGA-06-1802	0.84	0.139	0.775	0.0813	2	2	2	0.579	0.195	0	0	0.269	0.85	0.86	90	85

TCGA-06-1805	0.53	0.0238	0.125	0.759	3	2.8	2.89	0.0301	0.095	0.0948	0	0.48	0.53	0.58	80	90
TCGA-08-0244	0.9	0.0227	0.239	0.634	4	4.15	3.93	0.0817	0.157	0	0	0.80475	0.83	0.91	88	90
TCGA-08-0246	0.78	0.00455	0.722	0.876	2	2.8	2.62	0.64	0.0817	0	0.0431	0.76	0.77	0.78	80	90
TCGA-08-0344	0.91	0.0601	0.308	0.67	8	8.29	7.72	0.121	0.187	0	0	0.5885	0.68	0.84525	100	95
TCGA-08-0345	0.64	0.194	0.203	0.995	2	2.06	2.04	0.136	0.0671	0.0602	0.0173	0.11	0.62	0.65525	75	85
TCGA-08-0346	0.62	0.00697	0.386	0.975	4	5.48	4.16	0.274	0.112	0	0	0.44	0.45	0.46	85	90
TCGA-08-0347	0.76	0.00662	0.262	0.749	2	1.97	1.98	0.12	0.143	0.113	0.0162	0.75	0.76	0.77	80	80
TCGA-08-0348	0.87	0.0172	0.374	0.942	6	6.43	5.85	0.19	0.184	0	0	0.64	0.68	0.70525	95	90
TCGA-08-0349	0.81	0.00981	0.426	0.703	2	1.94	1.95	0.133	0.293	0.13	0.115	0.79475	0.82	0.83	75	85
TCGA-08-0350	0.82	0.0434	0.524	0.712	4	4.84	4.33	0.34	0.183	0	0.00185	0.64475	0.67	0.82	85	90
TCGA-08-0351	0.58	0.0624	0.24	0.97	4	5.6	4.09	0.183	0.0568	0	0.0322	0.41475	0.56	0.64525	90	80
TCGA-08-0352	0.63	0.0144	0.166	0.854	2	2.1	2.07	0.101	0.0655	0.0558	0	0.6	0.63	0.65	75	80
TCGA-08-0353	0.81	0.00634	0.289	0.695	2	2.02	2.02	0.126	0.162	0	0.0276	0.80475	0.82	0.83	70	85
TCGA-08-0354	0.82	0.0501	0.634	0.747	4	4.34	3.92	0.337	0.297	0	0	0.7	0.82	0.85	75	90
TCGA-08-0355	0.88	0.0163	0.352	0.682	4	4.26	3.99	0.175	0.177	0	0	0.75	0.79	0.81	80	90
TCGA-08-0356	0.73	0.041	0.243	0.898	4	4.87	4.1	0.159	0.084	0	0	0.57	0.57	0.73	50	85
TCGA-08-0357	0.87	0.00256	0.976	0.116	2	2.06	2.05	0.848	0.128	0	0	0.86	0.86	0.87	90	95
TCGA-08-0358	0.96	0.0194	0.213	0.838	4	4.04	3.96	0.0881	0.125	0	0.124	0.9	0.92	0.96	100	95
TCGA-08-0359	0.84	0.0141	0.583	0.694	4	4.13	3.79	0.177	0.406	0	0	0.83	0.84	0.85	70	90
TCGA-08-0360	0.55	0.00478	0.374	0.943	4	5.34	3.84	0.0871	0.287	0	0	0.55	0.56	0.56	75	85
TCGA-08-0375	0.99	0	0.592	0.396	4	4.32	4.3	0.353	0.239	0	0	0.99	0.99	0.99	80	100
TCGA-08-0380	0.4	0	0.86	0.763	2	2.63	2.25	0.443	0.416	0.0927	0.0088	0.4	0.4	0.4	100	100
TCGA-08-0389	0.99	0.0141	0.816	0.274	4	3.95	3.93	0.352	0.464	0	0	0.95	0.98	0.99	80	95
TCGA-08-0390	0.44	0.0838	0.734	0.965	4	6.7	4.07	0.372	0.362	0.000735	0.0756	0.26	0.43	0.46625	95	95
TCGA-08-0392	0.76	0.0414	0.895	0.905	6	6.74	5.6	0.271	0.624	0	0.00205	0.74	0.74	0.81	90	100
TCGA-12-0616	0.9	0.02	0.161	0.786	4	4.04	3.83	0.0144	0.146	0	0	0.84	0.85	0.92	75	100
TCGA-12-0618	0.93	0.163	0.214	0.991	4	4.09	3.94	0.0863	0.128	0	0.0715	0.42475	0.88	0.92525	97	100
TCGA-12-0619	0.74	0.0805	0.534	0.955	4	4.45	3.81	0.16	0.375	0	0.0702	0.55475	0.73	0.76575	95	100

TCGA-12-0620	0.69	0.0723	0.21	0.886	4	4.73	3.88	0.0558	0.155	0	0	0.52	0.665	0.68	85	100
TCGA-12-0654	0.66	0.0539	0.251	0.998	2	1.95	1.97	0.0842	0.167	0.167	0	0.55475	0.67	0.73575	80	95
TCGA-12-0656	0.66	0.0897	0.345	0.998	4	4.94	3.94	0.101	0.244	0	0	0.37	0.53	0.67	100	100
TCGA-12-0657	0.88	0.369	0.163	0.977	2	2.01	2.01	0.0949	0.0683	0.0654	0	0.07375	0.16	0.931	100	100
TCGA-12-0670	0.89	0.0164	0.35	0.988	2	1.95	1.96	0.102	0.248	0.117	0.128	0.85	0.88	0.91	100	100
TCGA-12-0688	0.8	0.155	0.161	0.992	2	2.01	2.01	0.0933	0.0674	0.0661	0.000229	0.12	0.81	0.83	75	100
TCGA-12-0692	0.88	0.126	0.277	0.652	4	4.21	3.95	0.0991	0.178	0	0.0765	0.44	0.79	0.86	90	100
TCGA-12-0703	0.76	0.195	0.181	0.999	2	2.02	2.01	0.102	0.0792	0.0621	0.0139	0.35	0.655	0.79	95	95
TCGA-12-0707	0.85	0.117	0.227	0.985	2	1.96	1.97	0.0994	0.128	0.125	0.000426	0.428	0.85	0.86	80	90
TCGA-12-0772	0.74	0.139	0.282	0.899	4	4.58	3.91	0.0727	0.21	0	0	0.32	0.735	0.76	95	100
TCGA-12-0773	0.82	0.258	0.19	1	2	2.14	2.12	0.126	0.0635	0	0.0399	0.36	0.85	0.93	80	100
TCGA-12-0775	0.46	0.0141	0.152	0.976	2	1.95	1.98	0.0559	0.096	0.0913	0	0.42	0.44	0.47	35	80
TCGA-12-0776	0.48	0.212	0.354	0.811	4	5.91	3.88	0.1	0.254	0	0	0.31	0.46	0.93525	70	90
TCGA-12-0778	0.65	0.141	0.379	0.774	4	5.39	4.2	0.284	0.0945	0	0	0.25	0.64	0.67	100	100
TCGA-12-0780	0.86	0.0974	0.206	0.991	4	4.26	3.95	0.0985	0.108	0	0	0.349	0.76	0.77	70	100
TCGA-12-0820	0.92	0.236	0.181	0.982	4	4.17	4	0.0994	0.0821	0	0	0.23	0.84	0.88	90	100
TCGA-12-0821	0.93	0.00874	0.333	0.777	2	1.86	1.87	0.133	0.201	0.196	0.00285	0.91	0.92	0.94	100	100
TCGA-12-0822	0.38	0.0734	0.393	0.993	4	7.39	4.05	0.242	0.152	0.00216	0.0602	0.14475	0.27	0.36	100	100
TCGA-12-0826	0.8	0.148	0.869	0.998	4	4.83	4.27	0.515	0.354	0	0	0.32475	0.65	0.76	80	100
TCGA-12-0827	0.58	0.108	0.69	0.948	6	8.44	5.74	0.158	0.532	0	0	0.16	0.26	0.42	100	100
TCGA-12-0828	0.79	0.253	0.121	0.975	2	2	2	0.0572	0.064	0.0534	0	0.18	0.35	0.87775	85	100
TCGA-12-0829	0.42	0.0141	0.214	0.989	2	2.01	2.01	0.112	0.102	0.0996	0	0.38	0.42	0.43	90	100
TCGA-12-1088	0.66	0.119	0.457	0.933	4	4.92	3.93	0.153	0.304	0	0	0.21425	0.52	0.67	85	85
TCGA-12-1089	0.84	0.107	0.2	0.969	4	4.28	3.92	0.0557	0.145	0	0	0.54	0.68	0.84	80	85
TCGA-12-1091	0.94	0.0049	0.227	0.953	2	1.82	1.83	0.0264	0.2	0.19	0	0.94	0.95	0.95	90	95
TCGA-12-1092	0.72	0	0.503	0.997	6	7.35	5.85	0.221	0.282	0	0	0.01	0.01	0.01	80	80
TCGA-12-1093	0.41	0.0927	0.276	0.998	4	6.68	3.92	0.0992	0.177	0	0	0.19475	0.26	0.59225	80	90
TCGA-12-1094	0.83	0.216	0.214	0.946	2	1.99	1.99	0.104	0.11	0.11	0	0.14	0.82	0.84	70	90

TCGA-12-1095	0.67	0.0856	0.211	0.973	4	4.92	3.95	0.0928	0.118	0	0	0.45	0.64	0.67	85	90
TCGA-12-1096	0.45	0.063	0.407	0.988	4	6.52	4.04	0.205	0.201	0	0.0367	0.24	0.45	0.48	85	90
TCGA-12-1097	0.88	0.0378	0.208	0.908	2	1.9	1.91	0.0573	0.15	0.133	0	0.8095	0.87	0.91	80	80
TCGA-12-1098	0.86	0.0424	0.334	0.851	4	4.48	4.13	0.202	0.133	0	0	0.74	0.79	0.85	95	90
TCGA-12-1099	0.77	0.0755	0.828	0.895	6	6.48	5.45	0.18	0.648	0	0	0.53	0.75	0.77	85	90
TCGA-12-1598	0.88	0.0954	0.492	0.618	4	4.06	3.81	0.134	0.358	0	0.157	0.46	0.82	0.95	85	95
TCGA-12-1599	0.74	0	0.149	0.984	2	1.99	1.99	0.0723	0.0762	0.0739	0	0.05	0.05	0.05	80	80
TCGA-12-1600	0.96	0.13	0.423	0.644	6	5.91	5.76	0.0985	0.324	0	0	0.5295	0.88	0.95575	80	80
TCGA-12-1601	0.81	0	0.0931	0.261	2	2	2	0.0251	0.068	0.00148	0	0.8	0.8	0.8	NA	NA
TCGA-12-1602	0.89	0.0825	0.693	0.784	4	3.45	3.29	0.0485	0.645	0.0185	0.0795	0.63	0.79	0.89525	60	80
TCGA-14-0736	0.73	0.0241	0.739	0.927	4	4.13	3.56	0.263	0.475	0	0	0.55	0.56	0.57	90	90
TCGA-14-0783	0.9	0.0689	0.514	0.834	6	6.32	5.89	0.205	0.309	0	0	0.6095	0.73	0.9	80	95
TCGA-14-0786	0.95	0.11	0.192	0.574	2	2	2	0.0845	0.108	0	0.0262	0.44475	0.94	0.96525	90	100
TCGA-14-0787	0.89	0.0148	0.247	0.967	2	1.91	1.92	0.0876	0.159	0.148	0.00517	0.85	0.9	0.91	85	95
TCGA-14-0789	0.52	0.191	0.141	0.992	2	2.01	2.01	0.0782	0.0626	0.0618	0	0.08	0.53	0.54	70	95
TCGA-14-0812	0.69	0.0487	0.257	0.491	2	1.89	1.93	0.00911	0.248	0.119	0	0.63	0.69	0.72	95	95
TCGA-14-0813	0.72	0.0998	0.411	0.775	4	4.64	3.9	0.101	0.31	0	0.0824	0.31475	0.56	0.75	65	95
TCGA-14-0817	0.73	0.119	0.658	0.764	6	7.12	5.74	0.258	0.4	0	0	0.44	0.73	0.76	80	95
TCGA-14-0865	0.99	0.0154	0.479	0.918	4	3.54	3.53	0.0334	0.446	0	0.175	0.93475	0.985	0.99	95	100
TCGA-14-0866	0.84	0.287	0.174	0.6	2	2.04	2.03	0.104	0.0691	0	0	0.13	0.82	0.92	95	100
TCGA-14-0867	0.88	0.117	0.712	0.892	6	6.22	5.71	0.277	0.434	0	0	0.53475	0.705	0.99	75	95
TCGA-14-0871	0.99	0.0391	0.992	0.737	6	5.54	5.5	0.304	0.689	0	0.00528	0.8655	0.99	0.99	95	100
TCGA-14-1034	0.75	0.187	0.333	0.998	4	4.55	3.91	0.132	0.202	0	0	0.18425	0.62	0.76	95	100
TCGA-14-1037	0.59	0.125	0.18	0.996	2	1.94	1.96	0.0557	0.124	0.124	0	0.19	0.58	0.6	85	100
TCGA-14-1396	0.81	0.0761	0.303	0.968	4	4.65	4.14	0.174	0.129	0	0	0.64	0.68	0.85	70	80
TCGA-14-1401	0.69	0.162	0.168	0.769	2	2.01	2	0.0924	0.0754	0.0607	0.0133	0.34	0.68	0.71	75	80
TCGA-14-1402	0.84	0.0159	0.222	0.686	2	2.01	2.01	0.0825	0.14	0.0771	0.0397	0.79	0.82	0.85	85	95
TCGA-14-1451	0.82	0.0117	0.205	0.901	2	1.93	1.94	0.0714	0.134	0.115	0	0.80475	0.83	0.85	75	100

TCGA-14-1452	0.82	0.146	0.179	0.849	2	2.05	2.04	0.114	0.065	0.0549	0	0.25	0.81	0.84	90	100
TCGA-14-1453	0.99	0.187	0.539	0.257	4	3.51	3.49	0.0904	0.449	0	0	0.56475	0.99	0.99	60	100
TCGA-14-1454	0.82	0.0896	0.886	0.71	4	4.95	4.42	0.606	0.28	0.00118	0	0.56	0.74	0.87	85	100
TCGA-14-1455	0.97	0.368	0.208	0.27	2	1.96	1.97	0.0723	0.136	0	0	0.08	0.97	0.99	85	100
TCGA-14-1458	0.41	0	0.241	1	4	6.82	3.97	0.0551	0.186	0	0	0.01	0.01	0.01	95	100
TCGA-14-1459	0.99	0.062	0.226	0.251	2	1.96	1.96	0.0578	0.168	0	0	0.99	0.99	0.99	60	100
TCGA-14-1794	0.48	0.0746	0.811	0.912	4	7.12	4.46	0.695	0.116	0	0	0.31475	0.46	0.6505	50	100
TCGA-14-1795	0.75	0.135	0.159	0.77	2	2	2	0.0905	0.0682	0.054	0.0132	0.19	0.76	0.76	80	100
TCGA-14-1821	0.77	0.0693	0.859	0.661	4	4.57	3.98	0.528	0.33	0	0	0.62	0.74	0.77	90	95
TCGA-14-1823	0.63	0.0967	0.781	0.552	2	2.61	2.39	0.548	0.233	0.000565	0	0.36	0.605	0.6605	90	95
TCGA-14-1825	0.92	0.143	0.904	0.329	4	4.39	4.2	0.395	0.51	0	0	0.23	0.81	0.88525	83	95
TCGA-14-1827	0.82	0.152	0.688	0.975	4	4.55	4.09	0.39	0.298	0	0	0.23475	0.57	0.83525	90	80
TCGA-14-1829	0.75	0.173	0.247	0.577	2	2.01	2.01	0.142	0.104	0.0698	0	0.2095	0.75	0.78	95	95
TCGA-15-0742	0.98	0.265	0.295	0.908	2	1.9	1.91	0.103	0.192	0.174	0	0.16	0.98	0.99	75	95
TCGA-15-1446	0.75	0.277	0.135	0.97	2	1.98	1.99	0.0607	0.0745	0.069	0.00173	0.05	0.75	0.75	95	95
TCGA-15-1447	0.77	0.113	0.692	0.954	2	2.66	2.51	0.557	0.134	0	0.111	0.308	0.76	0.80525	90	100
TCGA-15-1449	0.45	0.074	0.604	0.97	4	6.28	3.93	0.191	0.413	0	0	0.26	0.445	0.48	90	95
TCGA-16-0846	0.67	0.141	0.652	0.969	6	7.72	5.83	0.186	0.465	0	0	0.31	0.38	0.66	85	85
TCGA-16-0848	0.79	0.112	0.257	0.924	4	4.52	3.99	0.0799	0.177	0	0	0.36	0.65	0.75	95	95
TCGA-16-0849	0.94	0.177	0.0842	0.401	4	4.12	4	0.044	0.0401	0	0	0.27475	0.88	0.94	85	95
TCGA-16-0850	0.99	0.201	0.504	0.914	4	3.47	3.45	0.07	0.434	0	0.0823	0.34	0.97	0.99	95	98
TCGA-16-0861	0.83	0.0414	0.214	0.681	2	1.89	1.91	0.0643	0.149	0.102	0	0.73	0.81	0.87	80	90
TCGA-16-1045	0.62	0.059	0.436	0.727	4	4.7	3.68	0.0647	0.372	0	0	0.41475	0.45	0.64	95	95
TCGA-16-1047	0.97	0.0821	0.213	0.288	2	1.93	1.93	0.0731	0.14	0	0	0.9395	0.97	0.98	60	95
TCGA-16-1055	0.78	0.185	0.206	0.971	4	4.59	4.02	0.0747	0.131	0	0	0.06475	0.66	0.76	85	80
TCGA-16-1056	0.83	0	0.188	0.941	2	1.89	1.91	0.0273	0.161	0.124	0.0312	0.83	0.83	0.83	100	95
TCGA-16-1060	0.76	0.015	0.207	0.972	2	1.91	1.93	0.0323	0.175	0.12	0.0493	0.73	0.75	0.78	100	80
TCGA-16-1062	0.94	0.0109	0.157	0.62	2	2.01	2.01	0.0868	0.0706	0.0112	0.000811	0.91	0.94	0.95	90	90

TCGA-16-1063	0.9	0.0246	0.233	0.921	2	1.93	1.93	0.0709	0.162	0.142	0.0173	0.84475	0.9	0.93	100	95
TCGA-16-1460	0.75	0	0.171	0.864	2	2.05	2.04	0.162	0.00966	0.00293	0.1	0.05	0.05	0.05	90	80
TCGA-19-0955	0.48	0.168	0.239	1	2	2.09	2.04	0.162	0.0775	0.0534	0	0.07	0.44	0.56	95	90
TCGA-19-0957	0.84	0.0769	0.279	0.516	2	2.09	2.08	0.162	0.117	0	0	0.66	0.815	0.85	95	95
TCGA-19-0960	0.95	0.0128	0.209	0.999	2	1.94	1.94	0.0575	0.151	0.139	0.0126	0.92475	0.95	0.97	90	100
TCGA-19-0962	0.79	0.0093	0.19	0.991	2	1.92	1.94	0.0574	0.133	0.13	0	0.7695	0.8	0.8	100	100
TCGA-19-0963	0.74	0.189	0.196	0.798	2	1.95	1.97	0.0595	0.136	0.0979	0	0.21475	0.735	0.8	70	100
TCGA-19-0964	0.98	0.0131	0.105	0.822	2	1.91	1.91	0.00162	0.103	0.0707	0.0138	0.95	0.98	0.99	90	100
TCGA-19-1385	0.42	0.103	0.582	1	6	11.4	5.93	0.235	0.347	0	0	0.05	0.09	0.43	80	90
TCGA-19-1386	0.99	0.189	0.602	0.22	2	2.02	2.02	0.269	0.333	0	0	0.2	0.99	0.99	100	100
TCGA-19-1387	0.62	0.215	0.763	0.994	4	5.03	3.88	0.0629	0.7	0	0.0312	0.15	0.45	0.85	50	100
TCGA-19-1388	0.79	0.151	0.881	0.778	6	6.28	5.38	0.185	0.696	0	0	0.32	0.54	0.76	100	100
TCGA-19-1389	0.34	0.104	0.756	1	4	7.82	3.98	0.552	0.204	0	0.102	0.06	0.33	0.37	100	100
TCGA-19-1392	0.99	0.077	0.559	0.762	4	3.77	3.75	0.155	0.405	0	0.0796	0.70025	0.95	0.99	100	90
TCGA-19-1786	0.91	0.204	0.871	0.991	4	4.38	4.17	0.636	0.235	0	0	0.24475	0.82	0.87525	60	85
TCGA-19-1788	0.32	0	0.548	0.993	4	8.47	4.07	0.442	0.107	0	0	0.01	0.01	0.01	100	80
TCGA-19-1789	0.93	0.175	0.318	0.373	2	1.98	1.98	0.149	0.169	0.0665	0	0.14	0.93	0.94	90	90
TCGA-19-1791	0.99	0.166	0.853	0.116	2	1.99	1.99	0.363	0.49	0.00605	0	0.14475	0.99	0.99	95	95
TCGA-26-1438	0.42	0.0917	0.227	0.996	4	6.68	3.97	0.056	0.171	0	0	0.23	0.415	0.44	90	90
TCGA-26-1440	0.85	0.191	0.323	0.623	4	4.44	4.07	0.186	0.137	0	0	0.25	0.75	0.84	90	90
TCGA-26-1443	0.98	0.0937	0.13	0.42	2	2.03	2.03	0.0551	0.0752	0	0	0.93	0.98	0.98	80	85
TCGA-26-1799	0.89	0.374	0.736	0.188	2	2.04	2.03	0.469	0.267	0.0555	0.00259	0.09	0.885	0.90525	70	80
TCGA-27-1830	0.68	0.0863	0.6	0.391	2	1.89	1.92	0.276	0.325	0.127	0.0281	0.57475	0.68	0.75	70	85
TCGA-27-1832	0.58	0.0082	0.278	0.919	4	5.19	3.85	0.0475	0.23	0	0	0.55	0.56	0.58	95	95
TCGA-27-1833	0.99	0.271	0.196	0.49	2	2	2	0.0973	0.0991	0	0	0.17	0.99	0.99	95	95
TCGA-27-1834	0.77	0.114	0.605	0.8	6	6.72	5.63	0.179	0.426	0	0	0.41	0.58	0.76	65	85
TCGA-28-1745	0.4	0.0829	0.196	1	4	6.83	3.93	0.0551	0.141	0	0	0.08	0.24	0.27	90	90
TCGA-28-1746	0.95	0.139	0.617	0.186	2	2.09	2.09	0.232	0.385	0.00025	0	0.38475	0.95	0.99	90	90

TCGA-28-1749	0.99	0.331	0.164	0.598	2	2.02	2.02	0.0985	0.0657	0	0	0.16	0.99	0.99	80	90
TCGA-28-1750	0.67	0.0864	0.854	1	4	5.9	4.61	0.624	0.23	0	0	0.46475	0.55	0.67	80	90
TCGA-28-1751	0.43	0	0.198	1	2	2.05	2.02	0.127	0.0708	0.0708	0	0.05	0.05	0.05	75	90
TCGA-28-1752	0.62	0.137	0.285	0.999	4	5.1	3.92	0.0725	0.212	0	0	0.14	0.305	0.57	80	80
TCGA-28-1755	0.95	0.296	0.366	0.3	2	1.97	1.97	0.153	0.213	0.00653	0.0489	0.13	0.95	0.96	70	95
TCGA-28-1757	0.85	0.00601	0.165	0.342	2	2	2	0.0561	0.109	0.00136	0	0.83	0.84	0.85	85	90

1: Standard deviation of AGP values (std), obtained from 100 bootstrap runs.

2:Percent of genome changed (PC): the fraction of the genome with CNAs.

3:Percent of genome on-Point (PoP): the fraction of genome with CNAs near a canonical point at the optimal AGP.

4:Ploidy of the euploid population (2n/4n/6n).

5:Genomewide average ploidy of the aneuploid population (pl.tumor).

6:Average overall ploidy (pl.overall): the weighted average ploidy of the aneuploid and euploid portions.

7:Percent of genome amplified (%amp), the fraction of genome with copy number gains.

8:Percent of genome deleted (%del), the fraction of genome with copy number losses.

9:Percent of hemizygous deletion (%del.loh), the fraction of loss-of-heterozygosity segments due to single-copy loss.

10:Percent of copy neutral loss of heterozygosity (%cn.loh).

11-13:2.5%, 50% and 97.5% quantiles of the bootstrapped distribution of AGP.

14-15:Histopathologic report of tumor contents: percent of tumor cells and nuclei. These were obtained from TCGA, not derived from AGP inference.

Table 2.2: Selected molecular signatures distinguishing Typical (T) and Atypical (AT) GBMs

	Signatures*	N(AT)[1]	N(T)[1]	N(T-PN)[1]	Fold change (T/AT) Change(T/AT)[3]	P value[2]	Reference
Molecular	***IDH1* mut**	10	0	0			Nobusawa et al, 2009
	***EGFR* amp**	1	106	18		<2.2e-16	Kleihues et al, 1999
	***EGFR* OE**	1	50	5	3.44	3.99E-07	Kleihues et al, 1999
	***MDM2* OE**	1	18	3	1.65	3.11E-04	Kleihues et al, 1999
	***CDKN2A* del**	8	79	12		0.13	Kleihues et al, 1999
	***PTEN* mut**	3	32	6			Kleihues et al, 1999
	***FAS* OE**	2	44	3	2.4	2.86E-05	Tohma et al, 1998
	***TP53* mut**	12	33	8			Kleihues et al, 1999
	***PDGFRA* OE**	12	27	10	0.29	2.85E-04	Kleihues et al, 2007
	G-CIMP+	15	0	0			Cooper et al, 2010
Clinical	**Mean Age (Onset)**	38	58	58		4.63E-05	Kleihues et al, 1999
	Sex Ratio (M/F)	1.5	1.56	2.16			Kleihues et al, 1999
	Survival (days)	1,024	370	232		7.48E-07	Kleihues et al, 2007
	Necrosis (%)	6.75%	12.80%	15.7		9.28E-03	Kleihues et al, 2007

1:Counts of Atypical (AT), Typical (T) and Typical-Proneural (T-PN) samples with corresponding signatures passing a certain threshold. For CNA, the threshold is LRR ratio (base 2) greater than 0.2 or less than -0.2. For mutation, the counts are for the presence of validated non-silent mutations. For gene expression it is ±0.5 for logged (base 2) gene expression level when the median across the entire cohort is centered at 0 (therefore >0.5 is counted as OE).

2:P-values for comparing between Typical and Atypical samples, using the student T-test for expression and CNA, Age of Onset, and Necrosis, and the log-rank test for survival time.

3:Fold change (transformed to linear scale) of gene expression between T and AT groups.

*Abbreviations: mutation (mut), amplification (amp), deletion(del), overexpression (OE).

Table 2.3: Revised class assignment obtained in this work

GBM1	Classes	GBM2	Classes	Phillips et al.	Classes
TCGA-02-0001	Mes*	TCGA-02-0116	Mes	GSM96954	Classical
TCGA-02-0003	Prolif	TCGA-06-0137	Classical	GSM96963	Atypical
TCGA-02-0006	Mes	TCGA-06-0138	Prolif	GSM96984	Classical
TCGA-02-0007	Prolif	TCGA-06-0145	Classical	GSM96991	Atypical
TCGA-02-0009	Classical	TCGA-06-0148	Classical	GSM97014	Atypical
TCGA-02-0010	Atypical	TCGA-06-0154	Mes	GSM97048	Atypical
TCGA-02-0011	Atypical	TCGA-06-0155	Mes	GSM96965	Prolif
TCGA-02-0014	Atypical	TCGA-06-0156	Mes	GSM96973	Classical
TCGA-02-0015	Mes	TCGA-06-0169	Mes	GSM96972	Mes
TCGA-02-0016	Classical	TCGA-06-0176	Atypical	GSM96970	Prolif
TCGA-02-0021	Classical	TCGA-06-0192	Mes	GSM96969	Mes
TCGA-02-0023	Classical	TCGA-06-0201	Mes	GSM96966	Mes
TCGA-02-0024	Atypical	TCGA-06-0206	Mes	GSM96967	Mes
TCGA-02-0025	Mes	TCGA-06-0208	Classical	GSM97041	Prolif
TCGA-02-0026	Atypical	TCGA-06-0211	Classical	GSM97004	Mes
TCGA-02-0027	Classical	TCGA-06-0213	Mes	GSM97002	Mes
TCGA-02-0028	Atypical	TCGA-06-0216	Prolif	GSM96996	Prolif
TCGA-02-0033	Mes	TCGA-06-0649	Mes	GSM96992	Mes
TCGA-02-0034	Mes	TCGA-06-0686	Prolif	GSM96989	Mes
TCGA-02-0037	Mes	TCGA-06-0743	Classical	GSM96987	Mes
TCGA-02-0038	Classical	TCGA-06-0744	Classical	GSM96982	Mes
TCGA-02-0039	Mes	TCGA-06-0745	Prolif	GSM96981	Mes
TCGA-02-0043	Classical	TCGA-06-0747	Classical	GSM96980	Mes
TCGA-02-0046	Prolif	TCGA-06-0749	Prolif	GSM96964	Mes
TCGA-02-0047	Atypical	TCGA-06-0750	Mes	GSM96961	Mes
TCGA-02-0048	Prolif	TCGA-06-0875	Prolif	GSM96958	Mes
TCGA-02-0051	Mes	TCGA-06-0876	Classical	GSM96951	Mes
TCGA-02-0052	Prolif	TCGA-06-0877	Mes	GSM96952	Mes
TCGA-02-0054	Mes	TCGA-06-0878	Mes	GSM96955	Atypical
TCGA-02-0055	Mes	TCGA-06-0881	Mes	GSM96959	Prolif
TCGA-02-0057	Mes	TCGA-06-1084	Mes	GSM96995	Prolif
TCGA-02-0058	Atypical	TCGA-06-1086	Mes	GSM97009	Prolif
TCGA-02-0059	Mes	TCGA-06-1087	Prolif	GSM97008	Prolif
TCGA-02-0060	Atypical	TCGA-06-1800	Mes	GSM97000	Prolif
TCGA-02-0064	Mes	TCGA-06-1801	Prolif	GSM97011	Prolif
TCGA-02-0068	Mes	TCGA-06-1802	Mes	GSM97010	Prolif
TCGA-02-0069	Atypical	TCGA-06-1805	Atypical	GSM96977	Prolif
TCGA-02-0070	Mes	TCGA-12-0654	Mes	GSM97040	Prolif
TCGA-02-0071	Mes	TCGA-12-0656	Classical	GSM96953	Atypical

TCGA-02-0074	Prolif	TCGA-12-0657	Mes	GSM96962	Classical
TCGA-02-0075	Mes	TCGA-12-0670	Classical	GSM96978	Classical
TCGA-02-0079	Mes	TCGA-12-0688	Classical	GSM96979	Classical
TCGA-02-0080	Atypical	TCGA-12-0692	Classical	GSM96983	Classical
TCGA-02-0083	Classical	TCGA-12-0703	Classical	GSM96985	Classical
TCGA-02-0084	Atypical	TCGA-12-0707	Classical	GSM96988	Classical
TCGA-02-0085	Mes	TCGA-12-0772	Mes	GSM96990	Atypical
TCGA-02-0086	Mes	TCGA-12-0773	Atypical	GSM96994	Classical
TCGA-02-0089	Mes	TCGA-12-0775	Mes	GSM97007	Classical
TCGA-02-0099	Mes	TCGA-12-0776	Mes	GSM97018	Atypical
TCGA-02-0102	Classical	TCGA-12-0778	Mes	GSM97037	Atypical
TCGA-02-0104	Atypical	TCGA-12-0780	Classical	GSM97042	Atypical
TCGA-02-0107	Mes	TCGA-12-0820	Prolif	GSM96976	Classical
TCGA-02-0113	Classical	TCGA-12-0821	Prolif	GSM96974	Classical
TCGA-02-0114	Atypical	TCGA-12-0822	Mes	GSM96950	Mes
TCGA-02-0115	Classical	TCGA-12-0826	Classical	GSM96997	Classical
TCGA-06-0122	Mes	TCGA-12-0827	Atypical	GSM96993	Mes
TCGA-06-0124	Mes	TCGA-12-0828	Classical		
TCGA-06-0125	Classical	TCGA-12-0829	Mes		
TCGA-06-0126	Classical	TCGA-12-1088	Mes		
TCGA-06-0127	Classical	TCGA-12-1089	Classical		
TCGA-06-0128	Atypical	TCGA-12-1091	Classical		
TCGA-06-0129	Atypical	TCGA-12-1092	Mes		
TCGA-06-0130	Mes	TCGA-12-1093	Mes		
TCGA-06-0132	Mes	TCGA-12-1094	Mes		
TCGA-06-0137	Classical	TCGA-12-1095	Mes		
TCGA-06-0138	Prolif	TCGA-12-1096	Mes		
TCGA-06-0143	Mes	TCGA-12-1097	Prolif		
TCGA-06-0145	Classical	TCGA-12-1098	Classical		
TCGA-06-0147	Mes	TCGA-12-1099	Atypical		
TCGA-06-0148	Mes	TCGA-12-1598	Prolif		
TCGA-06-0152	Mes	TCGA-12-1599	Mes		
TCGA-06-0154	Mes	TCGA-12-1600	Classical		
TCGA-06-0156	Prolif	TCGA-12-1601	Mes		
TCGA-06-0157	Classical	TCGA-12-1602	Prolif		
TCGA-06-0158	Classical	TCGA-14-0736	Mes		
TCGA-06-0166	Prolif	TCGA-14-0783	Mes		
TCGA-06-0168	Mes	TCGA-14-0786	Classical		
TCGA-06-0171	Prolif	TCGA-14-0787	Classical		
TCGA-06-0173	Prolif	TCGA-14-0789	Mes		
TCGA-06-0174	Prolif	TCGA-14-0812	Mes		
TCGA-06-0176	Mes	TCGA-14-0813	Prolif		

TCGA-06-0184	Mes	TCGA-14-0817	Mes
TCGA-06-0185	Classical	TCGA-14-0865	Prolif
TCGA-06-0187	Mes	TCGA-14-0866	Classical
TCGA-06-0188	Prolif	TCGA-14-0867	Atypical
TCGA-06-0195	Prolif	TCGA-14-0871	Prolif
TCGA-06-0197	Mes	TCGA-14-1034	Mes
TCGA-06-0208	Classical	TCGA-14-1037	Mes
TCGA-06-0210	Mes	TCGA-14-1396	Mes
TCGA-06-0211	Classical	TCGA-14-1401	Prolif
TCGA-06-0214	Prolif	TCGA-14-1402	Classical
TCGA-06-0219	Prolif	TCGA-14-1451	Prolif
TCGA-06-0221	Atypical	TCGA-14-1452	Mes
TCGA-06-0237	Prolif	TCGA-14-1453	Classical
TCGA-06-0238	Prolif	TCGA-14-1454	Prolif
TCGA-06-0241	Prolif	TCGA-14-1455	Prolif
TCGA-06-0644	Mes	TCGA-14-1458	Atypical
TCGA-06-0645	Mes	TCGA-14-1459	Classical
TCGA-06-0646	Prolif	TCGA-14-1794	Prolif
TCGA-06-0648	Prolif	TCGA-14-1795	Prolif
TCGA-08-0244	Classical	TCGA-14-1821	Atypical
TCGA-08-0246	Mes	TCGA-14-1823	Mes
TCGA-08-0344	Atypical	TCGA-14-1825	Prolif
TCGA-08-0345	Mes	TCGA-14-1827	Classical
TCGA-08-0346	Mes	TCGA-14-1829	Mes
TCGA-08-0347	Prolif	TCGA-15-0742	Classical
TCGA-08-0348	Prolif	TCGA-15-1446	Classical
TCGA-08-0349	Prolif	TCGA-15-1447	Atypical
TCGA-08-0350	Atypical	TCGA-15-1449	Prolif
TCGA-08-0351	Atypical	TCGA-16-0846	Atypical
TCGA-08-0352	Mes	TCGA-16-0848	Prolif
TCGA-08-0353	Classical	TCGA-16-0849	Atypical
TCGA-08-0354	Mes	TCGA-16-0850	Atypical
TCGA-08-0355	Classical	TCGA-16-0861	Prolif
TCGA-08-0356	Mes	TCGA-16-1045	Mes
TCGA-08-0357	Classical	TCGA-16-1047	Classical
TCGA-08-0358	Classical	TCGA-16-1055	Mes
TCGA-08-0359	Prolif	TCGA-16-1056	Classical
TCGA-08-0360	Mes	TCGA-16-1060	Mes
TCGA-08-0375	Classical	TCGA-16-1062	Classical
TCGA-08-0380	Prolif	TCGA-16-1063	Classical
TCGA-08-0389	Prolif	TCGA-16-1460	Atypical
TCGA-08-0390	Mes	TCGA-19-0955	Mes

TCGA-08-0392	Mes	TCGA-19-0957	Prolif
TCGA-12-0616	Prolif	TCGA-19-0960	Atypical
TCGA-12-0618	Prolif	TCGA-19-0962	Mes
TCGA-12-0619	Mes	TCGA-19-0963	Prolif
TCGA-12-0620	Mes	TCGA-19-0964	Classical
		TCGA-19-1385	Mes
		TCGA-19-1386	Classical
		TCGA-19-1387	Prolif
		TCGA-19-1388	Mes
		TCGA-19-1389	Mes
		TCGA-19-1392	Prolif
		TCGA-19-1786	Classical
		TCGA-19-1788	Atypical
		TCGA-19-1789	Classical
		TCGA-19-1791	Classical
		TCGA-26-1438	Mes
		TCGA-26-1440	Classical
		TCGA-26-1443	Classical
		TCGA-26-1799	Prolif
		TCGA-27-1830	Mes
		TCGA-27-1832	Mes
		TCGA-27-1833	Classical
		TCGA-27-1834	Mes
		TCGA-28-1745	Mes
		TCGA-28-1746	Prolif
		TCGA-28-1749	Classical
		TCGA-28-1750	Mes
		TCGA-28-1751	Mes
		TCGA-28-1752	Mes
		TCGA-28-1755	Classical
		TCGA-28-1757	Classical

1:GBM sample assignments based on the revised classification system. Sample from the three cohorts were combined in the same table, where each row is not intended to show any sample matching between studies.
*Abbreviations: Proliferative (Prolif), Mesenchymal (Mes).

Table 2.4: Selected gene expression features distinguishing Typical GBM classes

		TCGA			Phillips		
Classes	**Genes**	**Classical**	**Prolif**	**Mes**	**Classical**	**Prolif**	**Mes**
Classical	***EGFR***	**2.6**	-0.35	-0.05	**1.17**	-1.97	-0.6
	CDKN2A	**-0.9**	0.35	-0.027	**-0.38**	1.5	0.32
Proliferative	***PCNA***	0.24	**0.4**	-0.17	0.45	**0.82**	-0.2
	TOP2A	-0.06	**0.99**	-0.64	0.27	**1.77**	-0.32
Mesenchymal	***CHI3L1***	0.52	-0.42	**1.2**	-0.052	-1.57	**1.36**
	TRADD	-0.047	-0.19	**0.26**	-0.029	-0.12	**0.22**
	RELB	-0.034	-0.26	**0.31**	-0.043	-0.16	**0.39**
	TNFRSF1A	0.14	-0.37	**0.57**	0.093	-0.57	**0.92**

Number in each entry is the average expression value (log 2 scale) for a given class for the gene indicated in the column Genes.

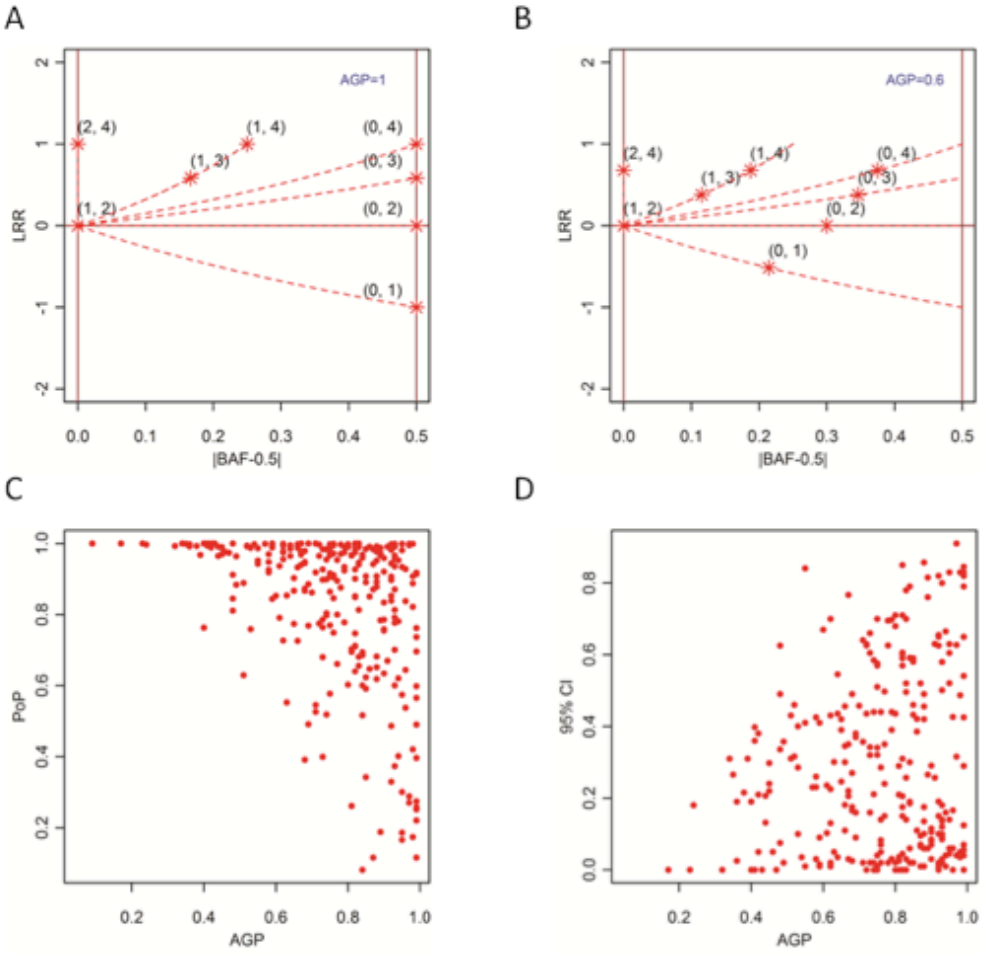
Table 2.5: Pairwise comparisons between GBM subtypes.

	PN/G-CIMP+	Prolif	Classical	MES
PN/G-CIMP+	-	1.5e-4	0.018	0.0015
Prolif		-	0.010	0.040
Classical			-	0.33
MES				-

	PN	NL	CL	MES
PN	-	0.059	0.046	0.091
NL		-	0.60	0.85
CL			-	0.92
MES				-

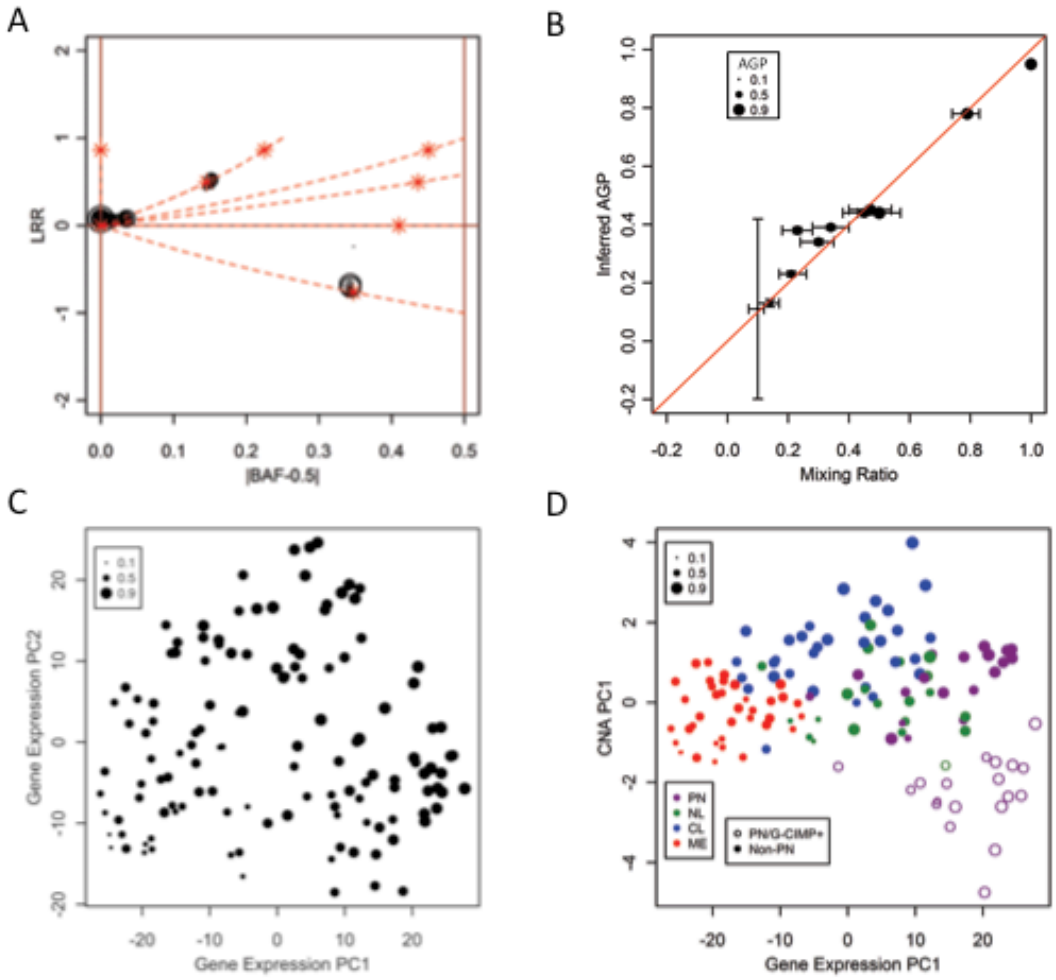
Log rank test was performed for each pair of subtypes compared. Upper table shows the results for revised GBM subtypes in this work, and lower table is for Verhaak's four subtypes.

Figure 2.1: Inference of Aneuploid Genome Proportion and its goodness-of-fit measures.



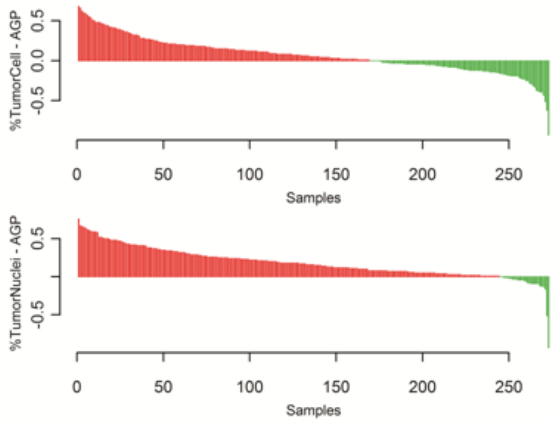
A. BAF-LRR plot for an idealized sample with 100% aneuploid cells. Canonical positions representing integer combinations of (N_B, N_A+N_B) are marked with red stars, with red dashed lines indicating the contraction paths when AGP is less than 100%. **B.** A hypothetic sample with AGP=0.6, with canonical points showing concerted contraction toward (1,2), the position of a normal diploid segment. **C.** AGP versus the PoP (Percent-on-Point), i.e., the fraction of CNA segments accounted for by canonical positions in the optimal mixing model. **D**. AGP versus the CI (Confidence Interval), defined as the range between the 2.5- and 97.5- percentiles in repeated runs of AGP estimates. Values were for GBM1 samples.

Figure 2.2: AGP and relationship to gene expression patterns A.



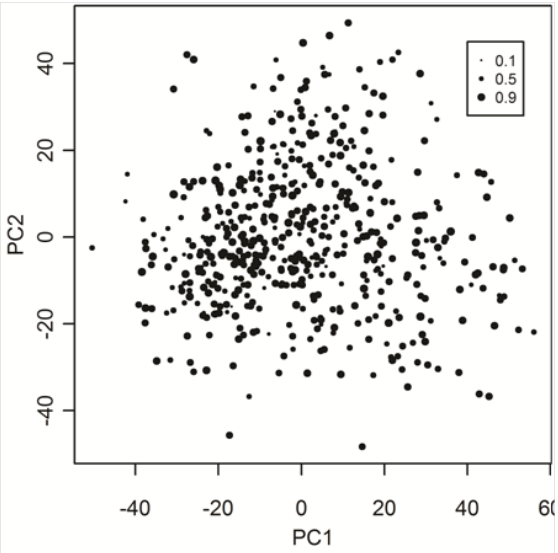
BAF-LRR plot of sample TCGA-02-0038 as an example of using allelic intensity data to estimate AGP. The x axis shows |BAF-0.5|, the absolute deviation of B allele frequency (BAF) between tumor and matched normal samples, at heterozygous SNP loci in the normal sample; y axis is the LRR, logR Ratio between tumor and normal samples. Canonical positions representing integer combinations of (N_B, N_A+N_B) are marked with red stars, with red dashed lines indicating the contraction paths when AGP < 1 (see also Figure S1). Most CNAs, shown as "bubbles", fell on canonical positions. The size of the bubble shows CNA length. PC (percent of genome changed) = 0.20 for this sample. Inferred AGP is 0.82. PoP (percent of changed genome on canonical points) = 0.99. **B.** Validation of AGP inference algorithm, using reference dataset GSE11976, for DNA pools of a breast cancer cell line mixed with a lymphoblastoid cell line at known ratios. Error bars show the 95% confidence intervals from the experimental procedures (horizontal) and from our bootstrap method (vertical). The red line has a slope of 1 and intercept of 0. **C.** Scatter plot of PC1-PC2 (the first two principal component scores) of GBM1 gene expression data. Symbol size is proportional to AGP as indicated in the legend. **D.** Scatter plot of PC1 of CNA (also shown on the x-axis in S4A) versus PC1 of gene expression data (shown on the x-axis in **1C**); Non- Proneural and Proneural/G-CIMP+ GBM samples were indicated by filled and open symbols, respectively.

Figure 2.3: Histolopathological estimates of tumor purities versus AGP.



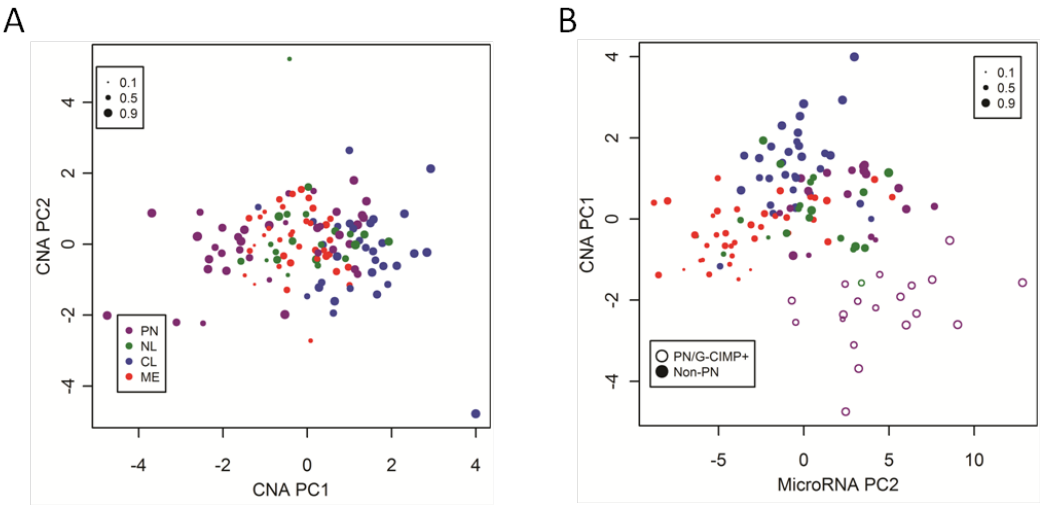
Comparison of AGP and clinically recorded "percent tumor cell"(upper panel) and "percent tumor nuclei" (lower panel), showing large deviations in many samples and generally higher estimates of tumor content in clinical records.

Figure 2.4: Relationship between AGP and gene expression pattern in ovarian cancer (OV).



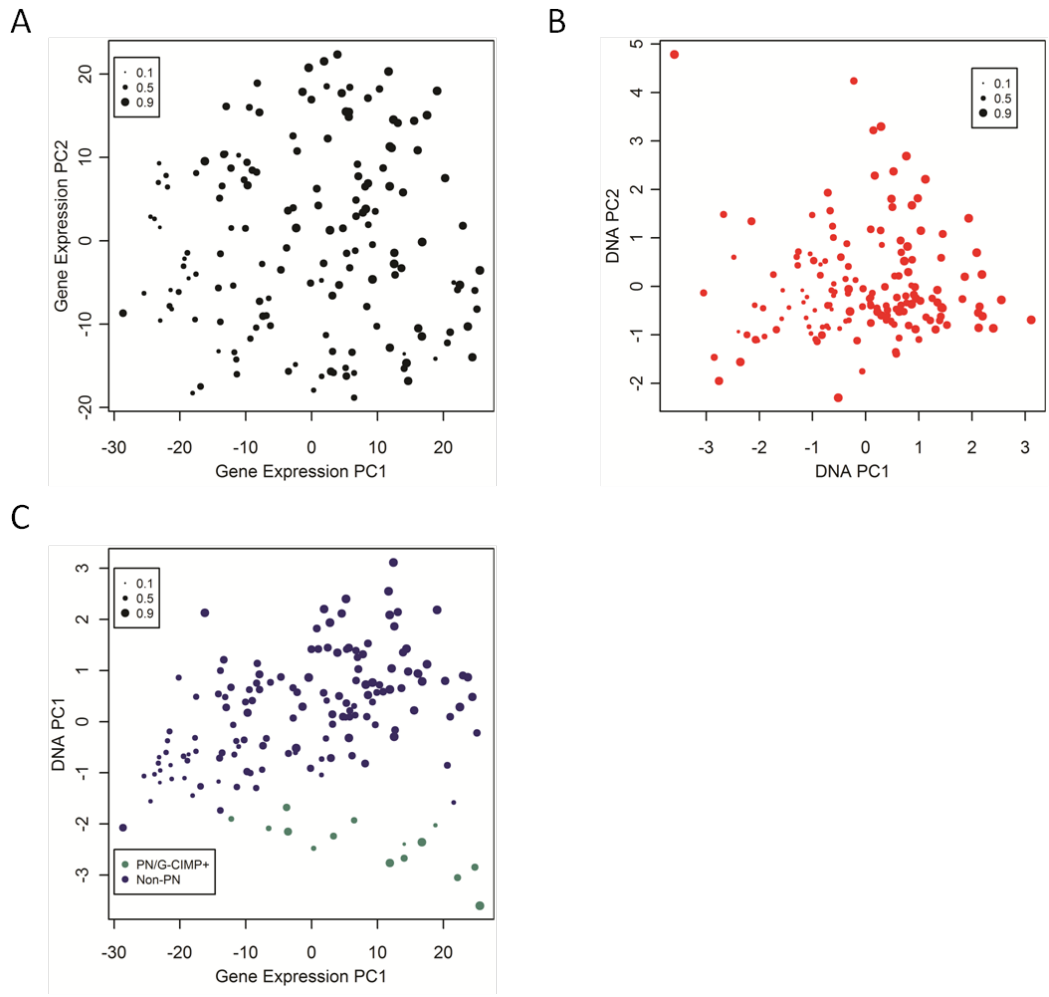
TCGA OV samples (n=504) were analyzed, and the gene expression PC1-PC2 plot showed an AGP gradient similar to that in Figure 1C.

Figure 2.5: PCA plots for CNA and MicroRNA joint analysis.



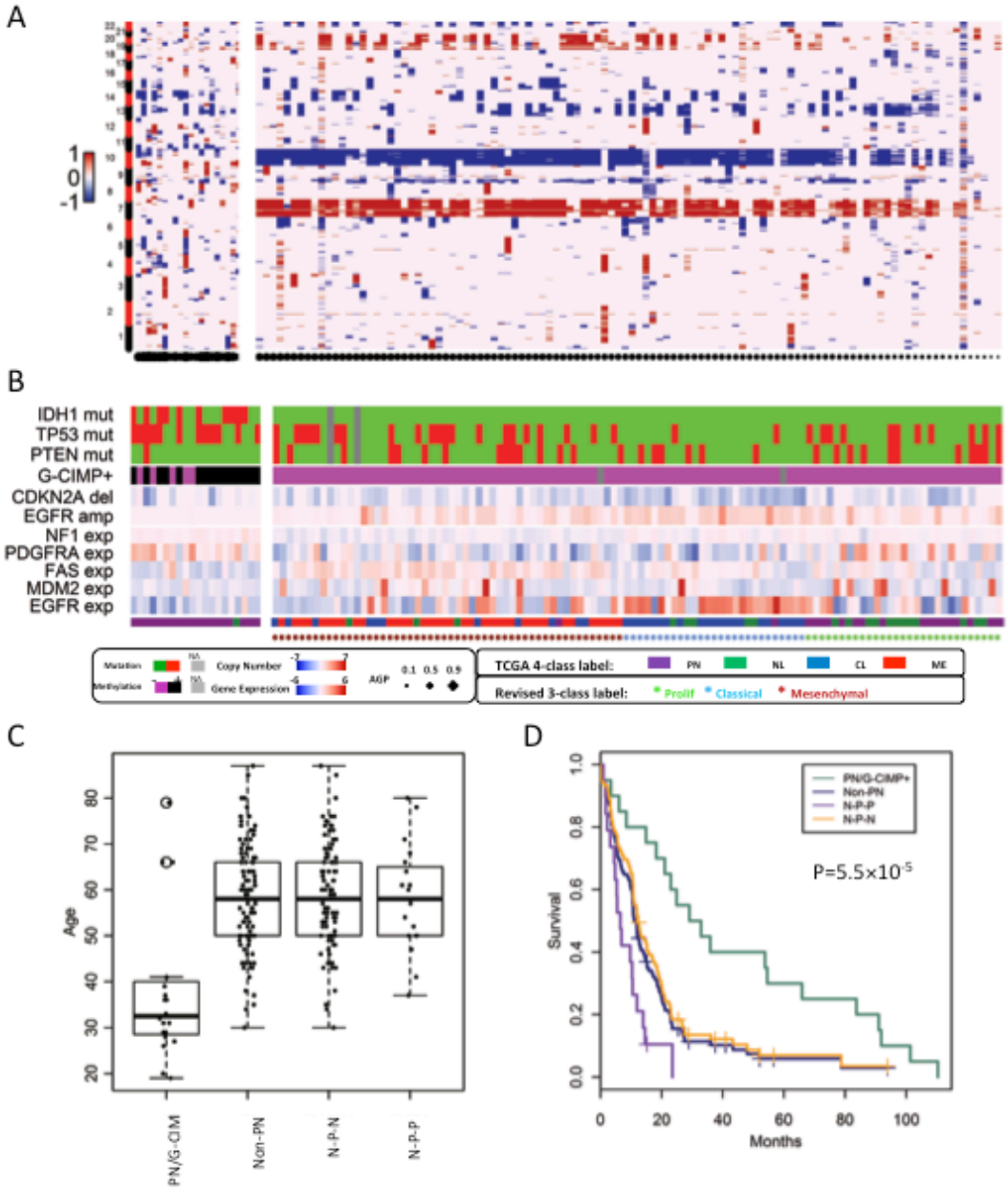
A. PC1-PC2 plot of average DNA copy number in each of 811 cytobands, with AGP indicated by bubble size and samples colored by the four-class assignment in Verhaak et al., showing the split of the Proneural subtype (purple). B. Scatter plot of PC1 of CNA (y) and PC2 of microRNA expression data (x).

Figure 2.6: Principal component analyses of gene expression and CNA data for GBM2.



A. PC1-PC2 plot for total DNA copy number data of 154 samples in GBM2, averaged in each of 811 cytobands, with AGP indicated by bubble size. B. PC1-PC2 plot for expression profiles in 1740 genes for the same samples in A. Patterns for both CNA and gene expression data were similar to the corresponding plots for GBM1 as shown in Figure 1C and Figure 2C. C. Joint use of DNA and gene expression data identified 15 PN/G-CIMP+ tumors in the GBM2 cohort, similar to the results for GBM1 as shown in Figure 2C-D.

Figure 2.7: Molecular and clinical features of Proneural/G-CIMP+ GBM A.



Heatmap of per-cytoband total copy number in Non-Proneural and Proneural/G-CIMP+ samples, with Chr1–22 arranged from bottom to top. Non-Proneural samples were ordered from left to right by decreasing AGP, and showed characteristic features, such as chr7 amplifications (shown in red) and chr10 deletions (in blue), across most samples, albeit with a gradient of magnitude. **B.** Selected molecular features, including, from top to bottom, presence or absence of non-silent mutations in *IDH1*, *TP53* and *PTEN* as reported by (2); G- CIMP+, a methylation signature described in (6); total copy number in *CDKN2A* and *EGFR*; expression levels of *NF1*, *PDGRFA*, *FAS*, *MDM2* and *EGFR*, as described in (2). The four classes defined in Verhaak et al., and the three classes defined in this work, are indicated as colored symbols in the bottom row. **C**. Distribution of age-of-diagnosis in Non-Proneural (n=110) and Proneural/G-CIMP+ (n=20) samples. Also shown are two subgroups of

Non- Proneural GBM: Proneural (N-P-P) and non-Proneural (N-P-N). **D.** Kaplan-Meier survival curves for Non-Proneural and Proneural/G-CIMP+ groups, with the latter showing better outcome (log rank test p-value=7.5E-7). The Non-Proneural group was further split into the former Proneural (N-P-P) and non-Proneural (N-P-N) samples.

Figure 2.8: Clustering pattern of three data types: PC1 of copy number data, PC1 of expression data, and PC2 of methylation data.

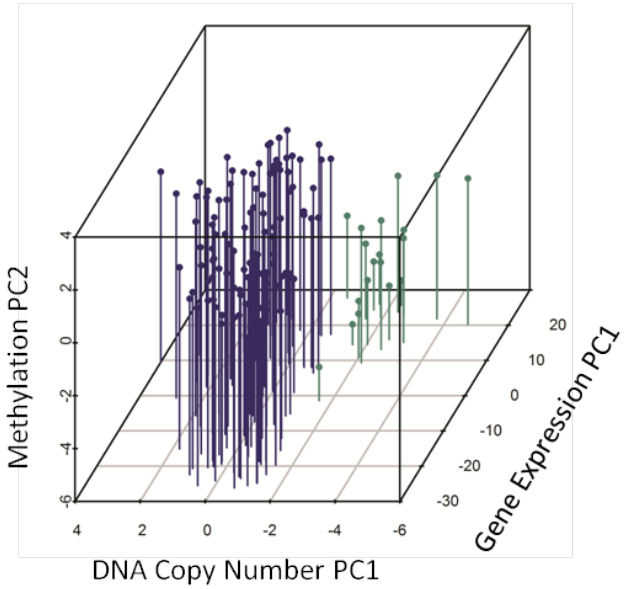
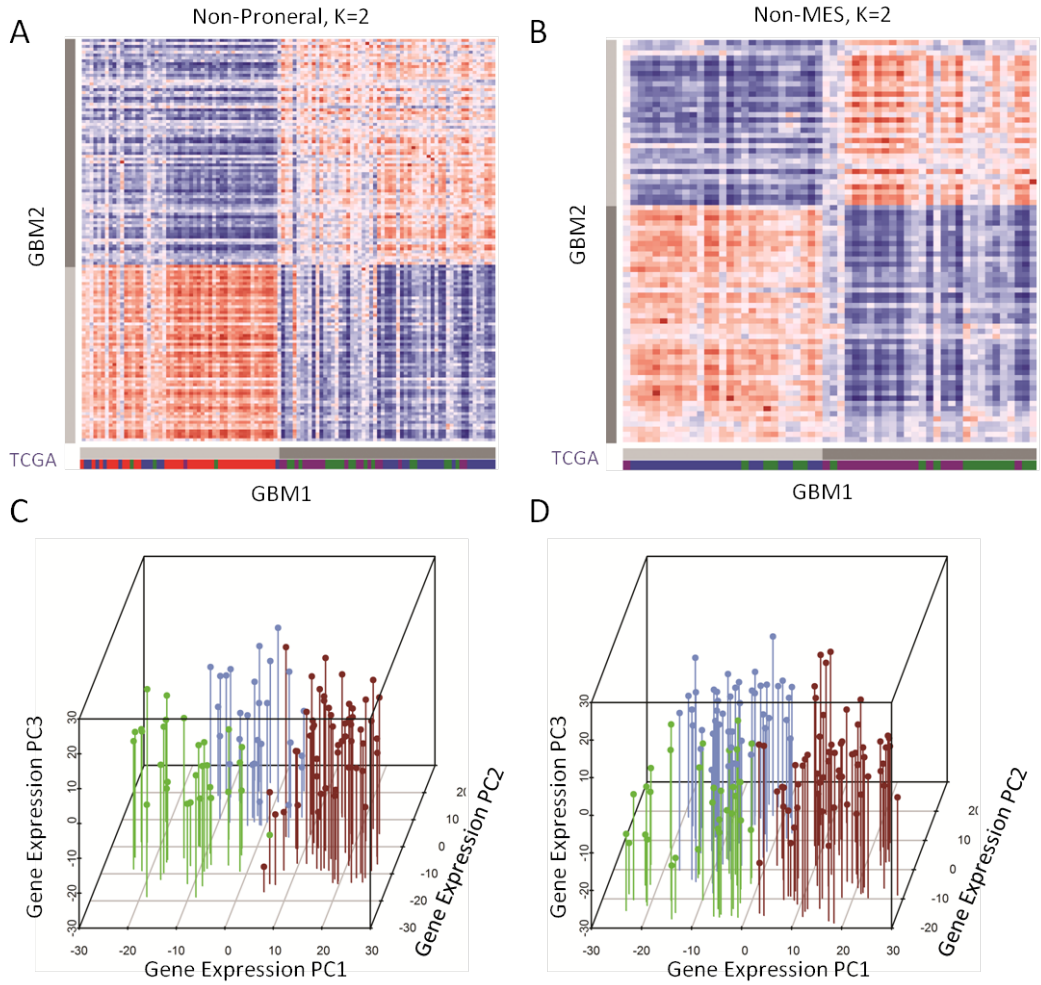
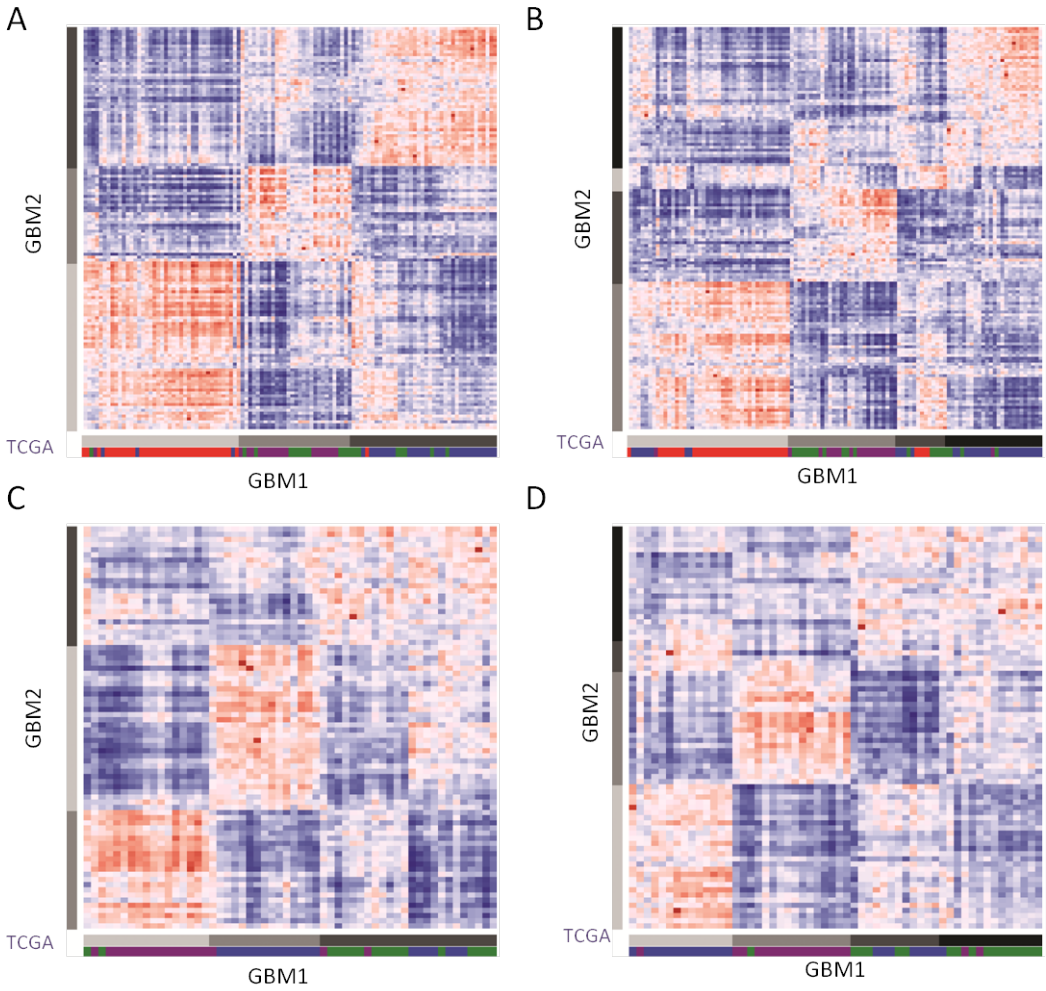


Figure 2.9: Classification of Non-Proneural GBM tumors.



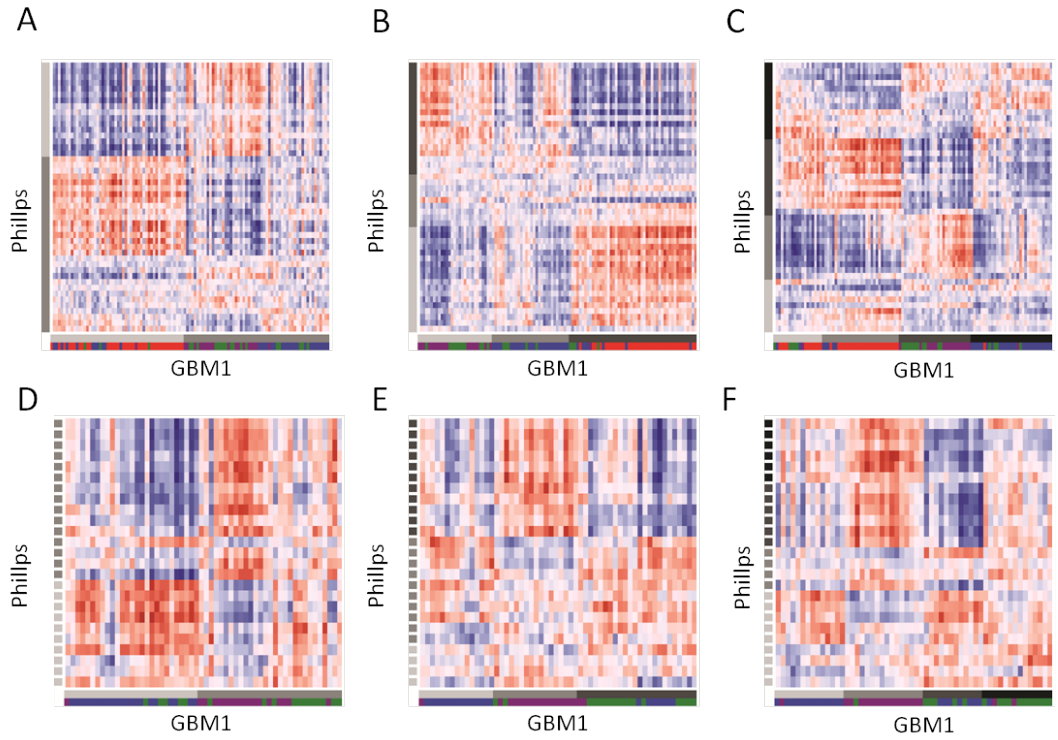
A-B. Heatmaps of the cross-correlation matrices between samples in GBM1 (left to right) and samples in GBM2 (top to bottom), for all Non-Proneural GBM tumors used in the first stage (A), and for one of the two subclasses discovered in the first stage (B). Consensus clustering was separately performed on both datasets with K = 2, with K = 3, 4 shown in Figure S8. Samples were ordered by the two-class assignment, as indicated by the two-color segments in the sidebar: vertical bar for GBM2, and horizontal bar for GBM1. The original four-class assignment from Verhaak et at. (2) was indicated in the four-color sidebar at the bottom. C-D. Three-dimensional PC1-3 plots of gene expression data for Non-Proneural samples in GBM1 (C) and GBM2 (D), showing coherent grouping of the three classes defined in A and B.

Figure 2.10: Cross-correlation analysis of GBM1-GBM2 at K=3 and 4. These plots complement Figure S5A-B, which showed k=2.



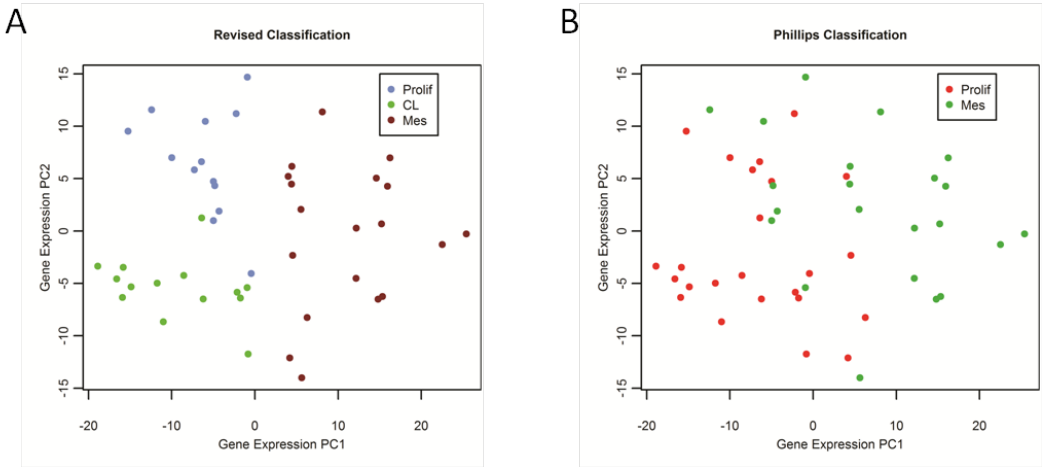
Shown are heatmaps of the cross-correlation matrices between samples in GBM1 (left to right) and samples in GBM2 (top to bottom), for all Non-Proneural GBM tumors used in the first stage at k=3 (A) and k=4 (B), and for the non-Mesenchymal class discovered in the k=2 first stage analysis, at k=3 (C) and k=4 (D). Samples were ordered by the three-class or four-class assignment, as indicated by the colored segments in the vertical bar for GBM2, and horizontal bar for GBM1. The original four-class assignment from Verhaak et at. was indicated in the four-color sidebar at the bottom. These plots revealed no clear one-to-one mapping between GBM1and GBM2 for either K=3 or K=4. C-D). The quality of mapping between batches can be quantified by the difference of average Pearson's r for diagonal and off-diagonal sample pairs, which are 0.428, 0.362 and 0.242 for K=2, 3, 4, respectively, for the first step analysis of all Non-Proneural tumors, and 0.442, 0.313 and 0.256 for K=2, 3, 4, respectively, for the second step analysis of Non-Mesenchymal samples.

Figure 2.11: Cross-correlation analysis between GBM1 and Phillips' dataset at K=2, 3 and 4



Correlation matrices for the first stage classification (**A, B, and C**, for k=2, 3, 4 respectively) and second stage (**D, E, and F**, for k=2, 3, 4 respectively), with procedures and sidebar labels similar to those shown in **Figure 10**

Figure 2.12: PCA plots for 46 Non-Proneural GBM samples in Phillips' dataset.



The same PC1-PC2 scatter plot, generated using 584 genes highly correlated with survival time, was shown in both **A** and **B**, and colored by the three-class assignment defined in this work (**A**) or the original Proliferative-Mesenchymal assignment by Philipps et al. (**B**), showing more coherent separation in **A**.

Figure 2.13: Comparison between the revised and the previous GBM classification systems.

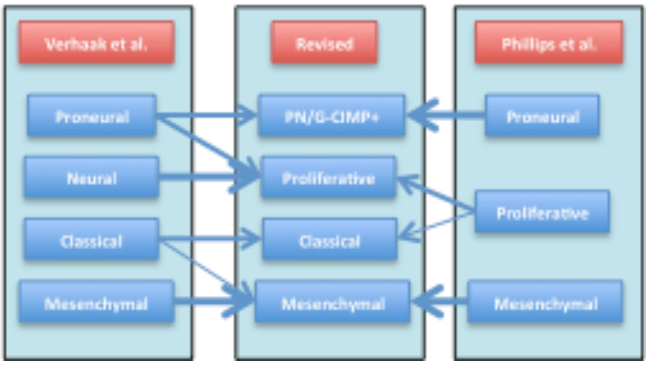
A

TCGA		Previous Classification			
		PN	NL	CL	ME
Revised	PN/G-CIMP+	19	1	0	0
	Prolif	17	12	0	0
	Classical	1	5	21	0
	MES	1	3	11	37

B

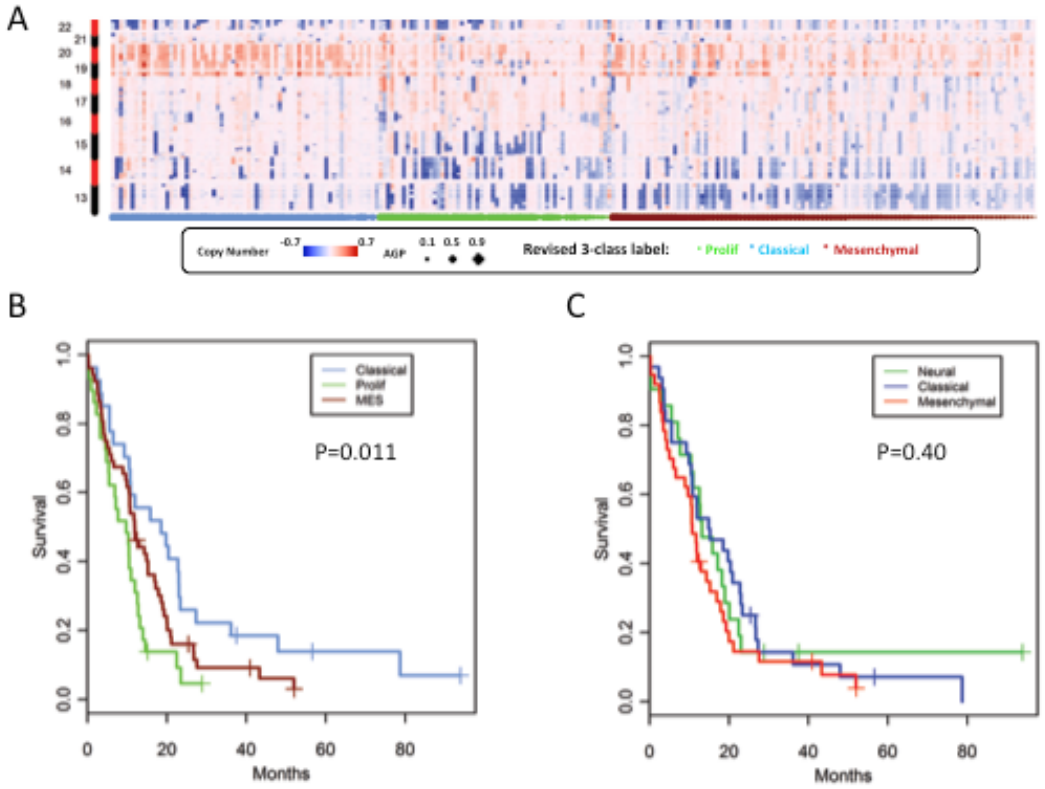
Phillips		Previous Classification		
		PN	Prolif.	MES
Revised	PN/G-CIMP+	10	0	0
	Prolif	0	13	1
	Classical	0	7	6
	MES	0	4	15

C



A. Cross-tabulation ("Confusion matrix") of samples between the current four-class assignment and that reported in Verhaak et al. (2) for TCGA samples. **B.** Cross-tabulation between the current four-class assignment and that reported in Phillips et al.(7). **C.** Correspondence of different classes across two datasets and the revised and previous classifications. Arrow width is proportional to the number of samples matched.

Figure 2.14: Molecular and clinical signatures for Non-Proneural GBM classes. A, Chr13–22 CNA patterns in the revised Non-Proneural GBM classes.



Class assignments were indicated by the stars at the bottom, with size proportional to AGP. B and C, Kaplan–Meier curves for GBM1 according to the revised classes (B) and the previous classes (C). The overall log-rank test for the 3 classes was significant in (B; P 1⁄4 0.011) but not in (C).

Figure 2.15: Survival time differences between GBM subtypes, compared between the current and previous classification systems.

Revised Classification				
	PN/G-CIMP+	Classical	Proliferative	MES
PN/G-CIMP+	-	9.5E-04	6.0E-07	4.7E-06
Classical		-	0.012	0.23
Prolif.			-	0.035
MES				-

Verhaak et al. Subtypes				
	PN	NL	CL	ME
PN	-	0.12	0.0044	0.065
NL		-	0.12	0.94
CL			-	0.18
ME				-

P values of pairwise survival time comparisons among the four classes defined in this work (upper table) and those defined in Verhaak et al. (lower table) were calculated from the log-rank test, showing greater differences in outcome among the revised classes.

Figure 2.16: Cox proportional hazard regression analysis with GBM subtypes as covariates, after adjusting age and Karnofsky performance scores (KPS).

A

Revised (n=128), PN-G-CIMP+ as reference		
	HR	p
Age	1.017	0.042
KPS	0.957	0.26
Classical	1.478	0.25
Mesenchymal	2.279	0.016
Proliferative	3.293	0.0018

B

Original (n=128), PN as reference		
	HR	p
Age	1.030	0.0001
KPS	0.925	0.043
Classical	0.898	0.68
Mesenchymal	1.118	0.67
Neural	0.738	0.33

C

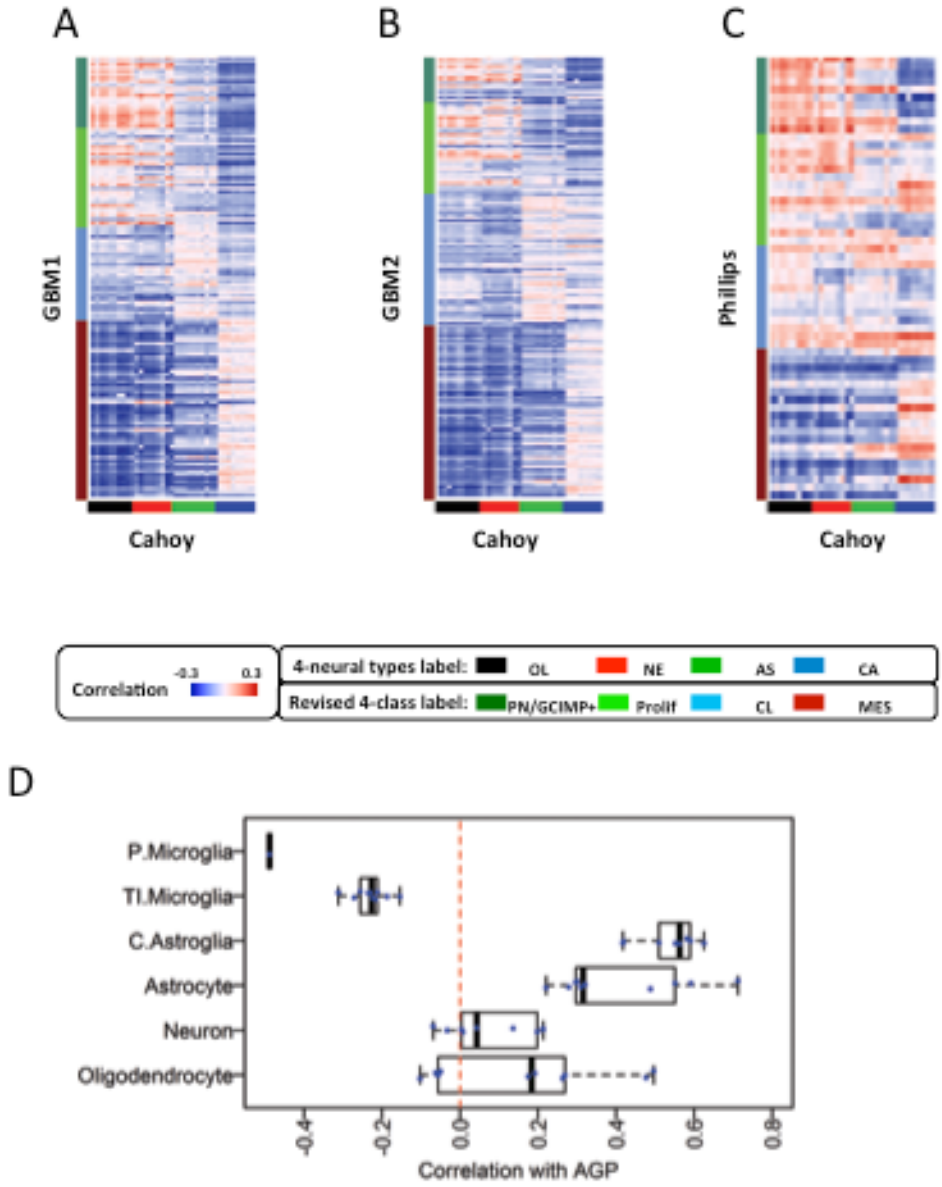
Revised (n=108), Classical as reference		
	HR	p
Age	1.025	0.013
KPS	0.946	0.17
Mesenchymal	1.393	0.22
Proliferative	1.958	0.029

D

Original (n=98), Classical as reference		
	HR	p
Age	1.037	0.000015
KPS	0.932	0.044
Mesenchymal	1.071	0.73
Neural	0.68	0.12

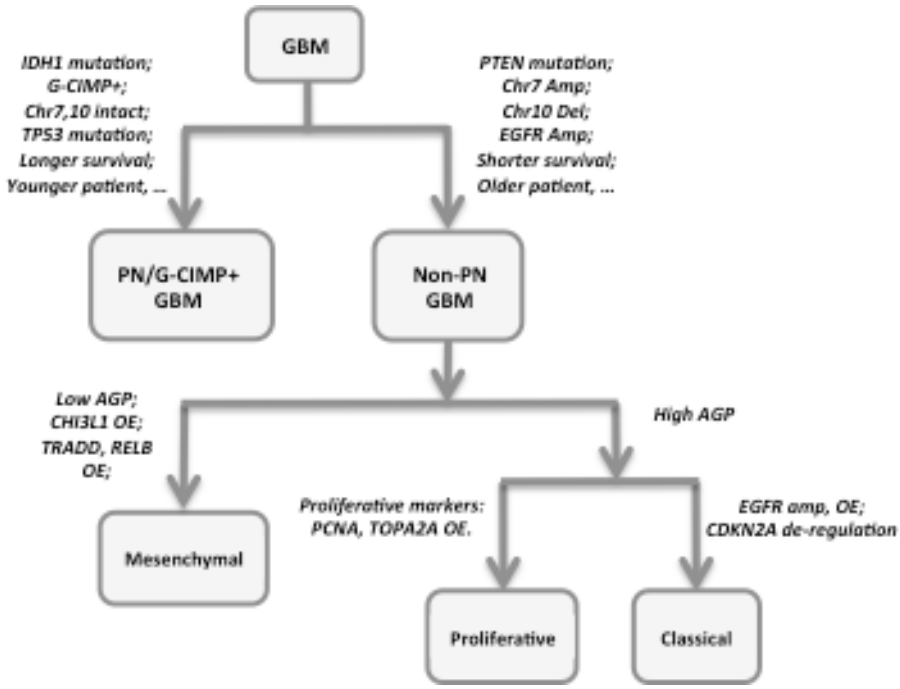
A) Hazard ratios (HR) and P-values (P) reported for Cox regression on 4 revised subtypes, 128 GBM1 samples, with Proneural/G-CIMP+ as reference category. B) Same analysis on 4 TCGA subtypes, with Proneural as reference. C) Cox regression performed on 3 revised subtypes, 108 GBM1 samples, excluding Proneural/G-CIMP+ and with Classical subtype as baseline. D) Same analysis on 3 TCGA subtypes, 98 GBM1 samples, excluding Proneural and with Classical as reference.

Figure 2.17: Inference of cell type composition of revised Non- Proneural classes.



A–C, heatmaps of the cross-correlation coefficients between the reference dataset of 38 samples of known neural cell types and samples in GBM1 (A), GBM2 (B), and Phillips' study (C). Colored segments in sidebars indicate sample assignments for 4 GBM classes or for the 4 neural cell types. D, distribution of the correlation coefficients between (1) AGP values of MES samples and (2) correlations with individual reference samples, for the 4 neuronal cell types in Cahoy and colleagues (GEO accession GSE9956) and 2 datasets for microglia (GSE25289 and GSE16119).

Figure 2.18: A proposed hierarchical classification scheme for GBM.



Joint use of DNA and mRNA data, along with patient age and outcome data, separates Proneural/G-CIMPþ GBMs from Non-Proneural GBMs in the first step of the decision tree. The 2 subsequent 2-way decisions define the 3 Non-Proneural classes, using features indicated in the diagram and the most informative transcripts in Supplementary **Table 4**.

2.13 Bibliography

Adamson, C., Kanu, O. O., Mehta, A. I., Di, C., Lin, N., Mattox, A. K. & Bigner, D. D. 2009. Glioblastoma multiforme: a review of where I have been and where I am going. *Expert Opin Investig Drugs,* 18**,** 1061-83.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X.*, et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature,* 403**,** 503-11.

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing, Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A.*, et al.* 2008. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci,* 28**,** 264-78.

Cairncross, J. G., Ueki, K., Zlatescu, M. C., Lisle, D. K., Finkelstein, D. M., Hammond, R. R., Silver, J. S., Stark, P. C., Macdonald, D. R., Ino, Y.*, et al.* 1998. Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *J Natl Cancer Inst,* 90**,** 1473-9.

Cooper, L. A., Gutman, D. A., Long, Q., Johnson, B. A., Cholleti, S. R., Kurc, T., Saltz, J. H., Brat, D. J. & Moreno, C. S. 2010. The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas. *PLoS One,* 5**,** e12548.

Ducray, F., Idbaih, A., De Reynies, A., Bieche, I., Thillet, J., Mokhtari, K., Lair, S., Marie, Y., Paris, S., Vidaud, M.*, et al.* 2008. Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile. *Mol Cancer,* 7**,** 41.

Edgar, R., Domrachev, M. & Lash, A. E. 2002a. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res,* 30**,** 207-10.

Edgar, R., Domrachev, M. & Lash, A. E. 2002b. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research,* 30**,** 207-210.

Guillemin, G. J. & Brew, B. J. 2004. Microglia, macrophages, perivascular macrophages, and pericytes: a review of function and identification. *Journal of Leukocyte Biology,* 75**,** 388-397.

Harrell, F. E., Jr., Lee, K. L. & Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med,* 15**,** 361-87.

Kleihues, P. & Ohgaki, H. 1999. Primary and secondary glioblastomas: from concept to clinical diagnosis. *Neuro Oncol,* 1**,** 44-51.

Mora, R., Dokic, I., Kees, T., Huber, C. M., Keitel, D., Geibig, R., Brugge, B., Zentgraf, H., Brady, N. R. & Regnier-Vigouroux, A. 2010. Sphingolipid Rheostat Alterations Related to Transformation Can Be Exploited for Specific Induction of Lysosomal Cell Death in Murine and Human Glioma. *Glia,* 58**,** 1364-1383.

Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W. L., Shay, T., Hamou, M. F., De Tribolet, N., Regli, L., Wick, W., Kouwenhoven, M. C. M.*, et al.* 2008. Stem cell-related "Self-Renewal" signature and high epidermal growth factor receptor expression associated with resistance to

concomitant chemoradiotherapy in glioblastoma. *Journal of Clinical Oncology,* 26**,** 3015-3024.

Murat, A., Migliavacca, E., Hussain, S. F., Heimberger, A. B., Desbaillets, I., Hamou, M. F., Ruegg, C., Stupp, R., Delorenzi, M. & Hegi, M. E. 2009. Modulation of Angiogenic and Inflammatory Response in Glioblastoma by Hypoxia. *Plos One,* 4.

Nobusawa, S., Watanabe, T., Kleihues, P. & Ohgaki, H. 2009. IDH1 Mutations as Molecular Signature and Predictive Factor of Secondary Glioblastomas. *Clinical Cancer Research,* 15**,** 6002-6007.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P.*, et al.* 2010. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell,* 17**,** 510-522.

Ohgaki, H. & Kleihues, P. 2007. Genetic pathways to primary and secondary glioblastoma. *American Journal of Pathology,* 170**,** 1445-1453.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics,* 5**,** 557-72.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J.*, et al.* 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res,* 16**,** 1136-48.

Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A.*, et al.* 2000. Molecular portraits of human breast tumours. *Nature,* 406**,** 747-52.

Phillips, H. S., Kharbanda, S., Chen, R. H., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L.*, et al.* 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell,* 9**,** 157-173.

Schwab, J. M., Frei, E., Klusman, I., Schnell, L., Schwab, M. E. & Schluesener, H. J. 2001. AIF-1 expression defines a proliferating and alert microglial/macrophage phenotype following spinal cord injury in rats. *Journal of Neuroimmunology,* 119**,** 214-222.

Shirahata, M., Iwao-Koizumi, K., Saito, S., Ueno, N., Oda, M., Hashimoto, N., Takahashi, J. A. & Kato, K. 2007. Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. *Clinical Cancer Research,* 13**,** 7341-7356.

Sun, L. X., Hui, A. M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R.*, et al.* 2006. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell,* 9**,** 287-300.

The Cancer Genome Atlas Research Network 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455**,** 1061-8.

The Cancer Genome Atlas Research Network 2011. Integrated genomic analyses of ovarian carcinoma. *Nature,* 474**,** 609-15.

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P.*, et al.* 2010. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* 17**,** 98-110.

Chapter 3. A general framework for analyzing tumor subclonality using DNA sequencing and SNP profiling data

3.1 Introduction

It have been recognized for nearly 40 years that cancer is a dynamic disease and its evolution follows a classical Darwinian process (Nowell, 1976, Fidler, 1978). After the proposal of the two-hit model of oncogenesis (Knudson, 1971), and especially after the discovery of multiple mutational milestones marking the linear progression from benign polyps to colorectal cancer (Fearon and Vogelstein, 1990, Vogelstein and Kinzler, 1993), it was briefly envisioned that cancer could be understood by simply finding the small number of events that act sequentially to drive step-wise clonal selection in most cancer cases. However, initial efforts to sequence most coding genes in tumor DNA revealed remarkable heterogeneity between tumors in each cancer type examined (Sjoblom et al., 2006, Wood et al., 2007, Jones et al., 2008): typically, very few (< 10) genes are mutated in >10% of tumors, but many (40-80) are mutated in 1-5% of tumors. Further, heterogeneity in cancer could manifest on other levels: not just among different patients, but also among tumors of different grades or organ sites in

the same patient, as well as among different cells within a tumor (Greaves and Maley, 2012, Yates and Campbell, 2012). Heterogeneity at any of these levels could confound diagnosis and treatment, and underlie the inherent evasiveness of this disease. Most genomic analyses to date, notably those led by the Cancer Genome Atlas (TCGA) Research Network (The Cancer Genome Atlas Research Network, 2008, The Cancer Genome Atlas Research Network, 2011, The Cancer Genome Atlas Research Network, 2012b, The Cancer Genome Atlas Research Network, 2012a) and the International Cancer Genome Consortium (ICGC) (Alexandrov et al., 2013) have focused on *inter*-tumor heterogeneity. These studies analyze hundreds of tumors per cancer type, relying on bulk tissue samples, usually from one tumor per patient. The data were primarily interpreted by regarding each tumor as a single population of cells with uniform character. Despite the inherent limitation of this assumption, as shown by the widely reported tumor-normal mixing (Van Loo et al., 2010, Li et al., 2012, Popova et al., 2009), large-scale inter-tumor comparisons have led to important new insights into significantly mutated genes (The Cancer Genome Atlas Research Network, 2008, The Cancer Genome Atlas Research Network, 2011), recurrently perturbed pathways, mutation signatures (Lawrence et al., 2013, Alexandrov et al., 2013), tumor subtypes (Verhaak et al., 2010a, Curtis et al., 2012), molecular predictors of outcome, and commonalities or distinctions among different cancer types (Garraway and Lander, 2013). However, these studies are not designed to adequately investigate intra-tumor heterogeneity. Ultimately, cancer genome evolution takes place at the single-cell level, and it is the cellular complexity and its dynamics that give rise to both intra- and inter-tumor heterogeneity. Currently,

cytogenetic methods are of low throughput and often cannot assure representative sampling. And the cost of single-cell sequencing (Navin et al., 2011, Shalek et al., 2013, Hou et al., 2012, Xu et al., 2012) remains prohibitively expensive for all but the proof-of-concept studies. Under such constraints, many groups have surveyed intra-tumor heterogeneity by comparing multiple specimens from the same patient by longitudinal sampling or spatial sampling (mainly for solid tumors). Almost invariably, analyses of longitudinal samples have uncovered dramatic temporal changes of the cancer cell population that often correlate with disease progression, severity, and treatment resistance (Keats et al., 2012, Ley et al., 2010, Durinck et al., 2011, Landau et al., 2013). Similarly, multi-region comparisons revealed extensive genomic variability across different geographic sectors of the tumor (Gerlinger et al., 2012, Sottoriva et al., 2013), or between the primary and metastatic tumors (Yachida et al., 2010). These studies, while using samples collected with a higher spatial or temporal resolution than those in TCGA and ICGC, often still contain heterogeneous populations of cells (Yachida et al., 2010, Campbell et al., 2010, McFadden et al., 2014).

Fortunately, when bulk tissue data describe the global average of multiple subpopulations of cells, it remains possible to statistically infer the number and genomic profile of such subpopulations. For example, when a sample is sequenced deeply, the somatic mutation frequencies sometimes cluster around a small number of distinct frequency "modes" (Nik-Zainal et al., 2012, Shah et al., 2012), suggesting that somatic mutations of similar frequencies may reside in the same population of cells and these cells may have descended from the same founder cell. For this reason, these mutations are said to belong to the same

"clone" or 'subclones", the latter referring to a clonal population of a relatively small cellular fraction. This inference task, essentially a deconvolution problem (or Blind Source Separation problem), presents many analytical challenges, since both the number of subclones and the genomic profile of each need to be estimated simultaneously, and somatic copy number alterations (sCNAs) and somatic single-nucleotide variants (SNVs) often reside in the same region yet have unknown phase or genealogical order. Currently available methods often need to invoke simplifying assumptions and often focus on a subset of the issues. For example, *ABSOLUTE* (Carter et al., 2012) uses sCNA data to estimate the global mixing ratio of aneuploid and euploid cells, but only under a tumor-normal, two-population assumption. When an sCNA or SNV is subclonal, *ABSOLUTE* makes the qualitative designation of "subclonal" without quantitatively estimating the clonality. Other types of compromises also accompany other methods, and I will defer the description of these limitations to the Discussion.

In this work, I developed Clonal Heterogeneity Analysis Tool (*CHAT*) as a general framework for estimating the cellular frequencies of both sCNAs and SNVs. It is suitable for analyzing genomewide SNP genotyping and DNA sequencing data for tumor-normal pairs (**Figure 3.1**). *CHAT* begins by identifying regions of sCNA or by partitioning the genome into bins; and for each sCNA or bin, it estimates a *local* mixing ratio, called segmental aneuploid genome proportion (sAGP), between a euploid population and a single aneuploid population carrying the local CNA. The assumption of *local* two-way mixing does not imply there are only two cell populations globally. It is akin to the infinite-site model in population

genetics, stating that each locus experienced only one copy number alteration, without a second overriding alteration or the reversal to the original germline state (i.e., back mutation). After calculating sAGP for every sCNA in the tumor, *CHAT* estimates the cellular prevalence of SNVs (also called cancer cell fraction, or CCF, as in (Landau et al., 2013)) by adjusting the observed somatic allele frequency (SAF) from sequencing data according to the background copy number status, while also considering the sCNA clonality (sAGP), the relative order of occurrence between the SNV and its associated sCNA, and their cis- or trans- relationship. Through simulation I show that *CHAT* performs well in quantitatively recovering sAGP, CCF, and the underlying evolutionary scenario. I have applied *CHAT* to calculate sAGP for sCNAs, and CCF for SNVs, across 732 human breast tumor samples previously analyzed for inter-tumor diversity by TCGA (The Cancer Genome Atlas Research Network, 2012b) (**Materials and Methods, Section 1**), and I will present two vignettes of the results. Lastly, I discuss the model identifiability issue and compare *CHAT* with several similar methods.

3.2 Data sources and sCNA identification

From the Cancer Genome Atlas Data Portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) I downloaded (1) the *Level-2* copy number data derived from the Affymetrix Genome-Wide Human SNP Array 6.0 (the "bi-allele" files) for 732 breast tumor DNA and their paired normal tissue DNA, and (2) the

VCF files for whole-exome sequencing data for a subset of 522 tumor-normal samples analyzed by TCGA (The Cancer Genome Atlas Research Network, 2012b). Of these, 445 samples have both SNP array and DNA sequencing available. The SNP array data were downloaded on 12/12/2012, while the sequencing data were downloaded on 3/22/2013. Each VCF file contains variant information for both the tumor and the paired normal sample. The procedures for variant calling and identification of somatic variants can be found in the Online Supplementary Methods of (The Cancer Genome Atlas Research Network, 2012b). Counts for somatic and reference alleles of both tumor and normal samples were extracted for use in this study.

In addition, I also downloaded the clinical annotation file, including the PAM50 designations of all the involved patients, on 12/17/2012.

sAGP estimation (see below) can be performed on two types of user-selected spatial units: (1) genomic bins, predefined for each sample, typically consisting of 500 heterozygous markers in the germline DNA, (2) naturally observed sCNA segments, which I detect using the Circular Binary Segmentation (CBS) method (Olshen et al., 2004), as follows. I independently perform segmentation on the LRR and the folded BAF (absolute value of BAF minus 0.5) values, using default parameters in the R package *DNAcopy* [46], except that "minimal markers required" was set to 5. With CBS results for both LRR and BAF, the two sets of change points are merged as follows: if a BAF change point falls within 5 markers of an LRR change point, either upstream or downstream, it is removed, i.e., only the LRR breakpoint is kept, under the assumption that the two change points capture the same event,

but the BAF change point is less accurately placed due to the greater sparsity of heterozygous markers.

After merging, the mean of LRR and folded BAF values are computed for each DNA segment (or the bin) in each sample, and used as input data for AGP and sAGP inference in the next step. For binned files, the bin length is on average 5.1Mb, and each sample has an average of 502 bins.

3.3 Inference of segmental aneuploidy genome proportion

3.3.1 Preview and hypothesis

The simplest form of intra-tumor heterogeneity is normal cell "contamination", i.e., mixture of aneuploid cells in the tumor with euploid cells in the surrounding normal tissue, the latter carrying the full and balanced set of chromosomes found in germline DNA. In our previous work (Li et al., 2012), I developed a method to calculate the overall fraction of the tumor cells, termed Aneuploid Genomic Proportion (AGP), assuming the global mixing of a tumor and a normal population. In brief, allelic intensity data from SNP genotyping arrays (or DNA sequencing) provide copy number information of the two parental chromosomes: n_a and n_b. Since n_a and n_b are both integers, the logarithm of total intensity ratio, LRR ~ $\log(n_a+n_b)$, and the observed B allele frequency, BAF = $n_b/(n_a+n_b)$, adopt a finite number of discrete BAF-

LRR combinations for different CNAs, and reside in "canonical positions" in the BAF-LRR plot. When aneuploid cells are mixed with euploid cells, logR-BAF positions of tumor sCNAs "contract" towards the euploid position; and different mixing ratios result in different degrees of contraction. Based on this feature I can quantitatively estimate a genomewide tumor mixing ratio (Li et al., 2012). Our algorithm relies on the same type of information, and shares the same goal, as several other methods (e.g., *ASCAT* and *ABSOLUTE*) (Van Loo et al., 2010, Carter et al., 2012). All of these methods assume that there is a single tumor population.

However, intra-tumor heterogeneity may also manifest as the co-existence of *multiple* tumor cell populations, each with its own copy number profile (Oesper et al., 2013). One example is shown in **Figure 3.2A**, where the sCNA segments marked in red show stronger contractions to the diploid track, for both LRR and BAF, than those marked in black; whereas those marked in green show even stronger contractions (**Figure 3.2A-B)**. As mentioned above, since all the sCNAs in black have similar cellular fraction values, I may infer the existence of a subclone, defined as a subpopulation of cells carrying the same set of events (the "black" sCNAs) due to their descent from a common ancestor tumor cell. This is the most parsimonious explanation why different somatic events in the genome could reach the same frequency. Meanwhile, another set of events, such as those in red, show a different cellular fraction values, suggesting the existence of a second subclone. When a tumor contains K tumor populations as well as a normal population, the term "purity" is no longer adequate as it requires K+1 mixing ratios to fully describe the tumor composition. Since the sCNA

segments with different mixing ratios are interspersed, this regional variation of clonality along the genome motivates us to extend the earlier concept of genomewide AGP to a new, segment-specific measure: sAGP.

The estimation of sAGP follows a similar approach as estimating the global AGP described in detail in Chapter 2, relying on the degree of contraction of each sCNA (**Figure 3.2B**). The method has the implicit assumption that at each sCNA the mixing involves only two populations, one of which is euploid. This assumption is largely satisfied when the somatic genome has experienced relatively sparse copy number changes, without global doubling or multiple rounds of complex local aberration. In effect, it assumes that, even though different sCNAs in the genome may belong to multiple populations of aneuploid cells, at each sCNA region there is only one aneuploid state that is mixed with the euploid state.

3.3.2. sAGP inference

As discussed in the main text, I jointly use BAF and LRR values to estimate sAGP for each sCNA, under the assumption of regional two-way mixing. The algorithm has three steps:

i. Data pre-processing

I assume the allele-specific copy number data are already in bi-allelic format, with the following fields in the input file: SNP ID, chromosome, position, A allele count, B allele count. To note, the allele counts may not be integer numbers, but could be real-numbered values from the original CEL file. SNP markers are first grouped into either bins or merged

sCNAs as described above. For each bin/sCNA, the median LRR and median folded BAF are calculated, and a segmentation file containing the above information for each segment is generated for each sample.

In the initial normalization of SNP array data the absolute LRR values depend on the genomewide average ploidy, which is affected by the relative abundance of different copy number states in the genome. For example, in a tumor with a high fraction of cells undergone genome-wide doubling, the DNA segment located near the origin of the BAF-LRR plot are AABB, instead of the normal diploid configuration AB, and the global ploidy can be well above 2. The first step of sAGP estimation is therefore to ascertain the genotype of the sCNAs near the origin, following the procedures described in (Li et al., 2012). This allows unambiguous assignment (when possible) of copy number states for other sCNAs in the genome and the calculation of average ploidy. The deviation of BAF and LRR values of the baseline sCNAs from (x_0,y_0) is also used to quantify $sd_{BAF}^2 \; and \; sd_{LRR}^2$ for use in downstream analysis.

ii. Estimate sAGP and absolute copy numbers

The method I used to estimate sAGP is extended from our AGP inference algorithm. For an sCNA with copy number configuration (n_b,n_t), where n_b is number of minor allele, and n_t number of total alleles, when mixed with a balanced diploid population its theoretical BAF and LRR values are:

$$BAF = \left| \frac{p \times n_b + 1 - p}{p \times n_t + 2 \times (1-p)} - 0.5 \right| + x_0$$

$$LRR = log_2\big(p \times n_t + 2 \times (1-p)\big) - 1 + y_0$$

where p is sAGP, and x_0, y_0 are the coordinates of the ($n_t = 2$, $n_b = 1$) state. When p changes, the points (BAF, LRR) follow a family of curved lines on the BAF-LRR plot, starting from the origin (x_0,y_0). Each line corresponds to a unique combination of (n_b, n_t) and is called a canonical line; and each point on this line uniquely corresponds to an sAGP value. The main task is to assign each observed segment to a canonical line. Due to noise, an sCNA does not locate precisely on a canonical line. Thus for each sCNA, I scan all possible canonical lines to find the one satisfying the following criteria:

(a) Distance to the closest canonical line $\leq 2*\sqrt{sd_{BAF}^2 + sd_{LRR}^2}$; where sd^2_{BAF}, and sd^2_{LRR} are the estimated standard errors of BAF and LRR values.

Sometimes multiple lines satisfy (a) and result in multiple sAGP and n_t estimates. In such cases I apply

(b) Choose $sAGP = argmin(F = n_t - 2\times ploidy + |p_s - p|)$; where p is sample-wide AGP and *ploidy* is the estimated global average ploidy from step ii). This criterion chooses the most probable canonical line as the one that results in a total copy number close to the genome-wide ploidy and an sAGP close to the global AGP.

If no canonical line can be found in (2), i.e., the deviation is greater than the specified 2X scale of the standard deviations of BAF and LRR markers, I consider the sCNA not meeting the regional two-way mixing hypothesis, and its sAGP is assigned NA, its n_b and n_t are also treated as missing values in downstream analysis.

3.4 Macroscopic clonal structure

3.4.1. Statistical modeling to infer macroscopic clonal structure

As explained above, sAGP values can be calculated for either predefined genomic bins or identified sCNAs. In the per-bin analysis, the user can choose to filter out the non-sCNA bins or those with very small sAGP values, as true sCNAs with length shorter than the bin width tend to have reduced sAGP estimates due to the flanking euploid regions. In our analysis of the breast tumor data I applied two filtering steps. First, I considered bins with median folded BAF ≤ 0.04 and absolute median LRR ≤ 0.16 to be euploid, and assigned sAGP = 0. Second, before sAGP clustering, I removed bins with sAGP ≤ 0.05 to remove the contribution of the small sAGP values. At this step there is an average of n = 224 bins left per sample. The two models of interest are evaluated in a maximal likelihood framework and the biological relevance of these models will be discussed in the next section.

For Model-1, the log likelihood has a uniform and a normal component:

$$l = \sum_{i=1}^{n} \ln\left(\frac{A}{range(\boldsymbol{Y})} + (1 - A) \times Norm(y_i, \mu, \sigma)\right)$$

where **Y** is the observed sAGP vector for a given sample, with components y_i , i=1,2,…,n, where n is the number of DNA segments after filtering. A is a scalar so that A/range provides the scaled uniform distribution. μ and σ are the mean and standard deviation of the single peak in the model following the normal distribution. I constrain A and μ in the range (0, 1).

The parameters A, μ and σ are estimated using the maximum likelihood approach, implemented in customized scripts (part of *CHAT*) written in the R statistical programming language.

Model-2 is fitted using a Dirichlet process Gaussian mixture model to infer the uncertain number of peaks and their relative abundances. The parameterization is as follows:

$y_i|\mu_i, \sigma_i \sim Norm(\mu_i, \sigma_i), i = 1,2 \ldots, n$

$\mu_i, \sigma_i|G \sim G$

$G|\alpha, G_0 \sim DP(\alpha G_0)$

$G_0 = Norm(\mu|\mu_1, \sigma/k_0)InvWi\square hart(\sigma|\nu_1, \psi_1)$

$\alpha|a_0, b_0 \sim \Gamma(a_0, b_0)$

$k_0|\tau_1, \tau_2 \sim \Gamma(\tau_1/2, \tau_2/2)$

Together these expressions describe a standard Dirichlet process mixture of normal model (Escobar and West, 1995). The implementation of the MCMC fitting is via R package *DPpackage* (Jara et al., 2011). There are different ways to specify the prior parameters for the normal mixture model. The baseline Gaussian distribution G_0 relies on three prior parameters, μ_1, σ and k_0, where σ is explicitly modeled by an Inversed Wishart distribution with priors ν_1 and ψ_1, and k_0 follows a Gamma distribution. In practice, the hyperpriors, ν_1, ψ_1 and k_0 can also be allowed to be random variables with a given prior distribution, and the model will have higher power to fit minor peaks in the data. In this work I used a conservative setting of prior parameters in terms of peak discovery sensitivity.

Model-1 cannot be included as special case of Model-2, since when y is truly uniformly

distributed, Dirichlet process tends to call multiple peaks instead of one peak, even with current conservative prior setting. Our solution is to fit both models, then numerically compute the likelihood of each model, and use Bayesian Information Criterion (BIC) to select the better model. Model-1 has three free parameters: A, μ and σ, while Model-2 has seven: a_0, b_0, k_0, ν_1, ψ_1, τ_1 and τ_2.

3.4.2 Evolutionary interpretations of statistical models

When there are a sufficient number of sCNAs or bins covered by sCNA, *CHAT* produces a sufficient number of sAGP values; and their distribution could inform the clonal structure of the tumor. First, for some tumors the sAGP histogram may contain a single peak, potentially accompanied by a flat (nearly uniform) background distribution (Model-1). This pattern can arise in a tumor containing a single clone that cover a large fraction of the sCNA-bearing portion of the genome, potentially with many other clones that cover much smaller portions of the genome and they are undiscernible in the sAGP spectrum. Second, for other tumors the histogram may follow a multi-modal distribution (Model-2), representing a number of distinct clones, each with a different sAGP, and each covering a comparable portion of the genome as to be recognizable in the histogram (an example is shown in **Figure 3.2C**).
In all, there are three attributes of each sAGP histogram. (1) The *number* of the modes corresponds to the number of identifiable cell populations, each with a different sAGP value. (2) The *positions* of the modes denote the clonality of each cell population. The right-most

peak represents the population with the highest sAGP, and is typically called the *dominant clone*. The peaks to the left of the dominant clones are often called *subclone 1*, *subclone 2*, etc. (3) The *areas* under the peaks reflect the number of the sCNAs, or the regularly spaced bins, that belong to each cell population. Note that the right-most peak may not have the largest area, thus the dominant clone may not cover the widest portion of the genome.

There are at least two ways to define the spatial unit in the sAGP analysis, and *CHAT* provides both options. The first is to calculate sAGP for regularly spaced bins, either for a fixed window width or for a fixed number of SNPs. The resulting sAGP values resemble the conventional genetic "markers"; and each tumor has a guaranteed number and density of such markers to construct the sAGP histogram, which is interpreted analogously to the allele frequency spectrum in population genetics studies. However, the bins don't match the naturally occurring sCNAs, which are highly variable in lengths, from tens of kb to entire chromosome arms. The sCNAs shorter than the bin width would have their true sAGP values "diluted" by flanking euploid segments in the same bin; whereas those longer than the bin width would generated a string of correlated sAGP values as the same sCNA is artificially divided into multiple adjacent bins, thus violating the assumption that sAGPs are independent. In the second option, *CHAT* will apply the identified sCNA as the naturally occurring spatial unit for sAGP calculation. While this has the advantage that all sAGPs are truly independent, there are two disadvantages. First, the longer (or shorter) sCNAs provide more (or less) precise estimates of sAGP, but this information of confidence was discarded, as it is also the case in (Oesper et al., 2013). Two, there will be large tumor-tumor variations in the number

of sCNAs, and some tumors may not have enough sCNAs to construct an informative histogram for estimating clonal composition. In short, the per-bin sAGPs (option 1) are derived from segments of similar length and have similar confidence intervals—they are identically distributed but not independent random variables. Conversely, the per-sCNA sAGPs (option 2) are independent, but are not identically distributed due to varying lengths. Rigorously speaking, neither is suited for analyzing macroscopic clonal architecture but can be applied in exploratory analysis, especially when there is no other data type such as the SNVs (see below).

When the primary goal of using *CHAT* is to accurately estimate CCF, which relies on accurate sAGP values, the user is advised to calculate sAGP using sCNAs as the unit rather than the bins. Alternatively, when the primary goal is to explore clonal composition of a tumor, and if there are too few sCNAs and if most of them are very large, it is beneficial to increase the number of informative features, just as the detection of population stratification requires many ancestry informative markers. Here the user may choose regularly spaced bins to increase the number of available sAGPs. In fact, when sCNAs are few and large, it is more advisable to collect sequencing data; and if the mutation rate is high and/or the entire genome is sequenced (as opposed to small targeted regions), it is better to rely on the CCF histogram to estimate clonal structure. CCF distributions have the important advantage of meeting the condition of independent and identically distributed variable. Ultimately, the best approach is to integrate the sAGP and CCF distributions in estimating clonal structure.

3.5 Estimating cell fractions of somatic mutations

3.5.1 Nature of the problem

The next step of *CHAT* turns from estimating sAGP of sCNAs to estimating the frequency of cells carrying a specific *mutation*, i.e., single nucleotide variant (SNV) or small insertion/deletion (indel). Here the method addresses the case where the tumor DNA has been sequenced, either for the whole genome or for a targeted subset, such as the exome. The input of the analysis is the observed number of reads in the sequence data containing the mutation as well as those containing the un-mutated allele. The relative fraction of mutation-bearing reads is termed *somatic allele frequency*, or SAF. Following (Landau et al., 2013), I adopt CCF to denote the percentage of cells in the tumor sample carrying a specific somatic mutation. CCF is also termed *cellular prevalence* in (Roth et al., 2014). The goal is to use the observed SAF to estimate the unknown CCF.

If the mutation resides in a normal diploid region, it typically occurs on the background of one of the two parental chromosomes, contributing to about half the sequence reads in this region. In this simple case, if the fraction of cells carrying the mutation is CCF, the expected fraction of sequence reads carrying the mutation, SAF, is simply a binomial variable with an expected value of CCF/2. I therefore can estimate CCF by SAF times 2. However, if the mutation resides in an sCNA, the relationship between CCF and SAF depends on the copy

number configuration: copy neutral loss-of-heterozygosity (CN-LOH), deletion, amplification, etc.) and its sAGP. Further, it also depends on the chromosomal background in which the mutation occurs. For example, in a region of heterozygous amplification where one of the chromosomes has been duplicated, if the mutation occurs on the duplicated chromosome, it will contribute twice the number of sequence reads than the case where it occurs on the un-duplicated chromosome. Lastly, if the mutation occurs after the duplication has happened and the duplication-bearing clone is undergoing expansion, only a subset of the duplication-bearing cells will carry the mutation, and the relative size of this subpopulation can be any value in 0-100% and will also affect the relationship between CCF and SAF. In this following I systematically consider these possible scenarios. I will make the parsimonious assumption that each mutation only occurred once in the evolutionary history of the tumor cell population, therefore I will ignore the possibility of recurrent mutation at the same position, or simultaneous emergence of the same mutation is different subpopulations of cells. I will treat SNVs and indels equivalently, and use the term "mutation" to denote both.

3.5.2 Order-phase scenarios between sCNA and SNV

For a somatic mutation revealed by tumor DNA sequencing, with an observed SAF value, I consider the task of estimating CCF if this mutation resides in an sCNA, and the sCNA has been discovered by either SNP array genotyping data (Van Loo et al., 2010, Carter et al.,

2012) or by sequencing data (Nik-Zainal et al., 2012, Landau et al., 2013). I assume that the sCNA has been well characterized, such that I already know n_a and n_b, the copy number of its major and minor alleles, respectively, i.e., $n_a \geq n_b$, and $n_t = n_a + n_b$ is the total copy number. I also assume that its sAGP has been calculated using the method described above, and that SAF has been corrected for known sequencing errors and local biases (Lawrence et al., 2013, Cibulskis et al., 2013). Below I present the CCF estimation procedure for the case of heterozygous amplification ($n_a = 2$, $n_b = 1$).

When a mutation resides in an sCNA region, there are three main scenarios that describe the possible mutation-sCNA combinations in terms of their relative temporal order and the chromosomal background of the mutation (**Figure 3.3**):

A) The mutation and sCNA emerged sequentially, with the mutation occurring first, and the sCNA occurring in a subset of mutation-bearing cells (**Figure 3.3A**). This led to the co-existence of three subpopulations: the original euploid mutation-free cells, with the population fraction of r_0; cells carrying the mutation only, with a fraction of r_1; and cells carrying both the mutation and the sCNA (r_2). The last subpopulation has two alternative outcomes: $\mathbf{A_1}$: the duplication occurred on the mutation-bearing chromosome, and $\mathbf{A_2}$: the duplication occurred on the mutation-free chromosome. Intuitively, $\mathbf{A_1}$ will have higher SAF than $\mathbf{A_2}$ with the same (r_0, r_1, r_2) fractions.

B) Like **A**, the mutation and sCNA emerged sequentially; but unlike **A**, the sCNA occurred first, with the mutation occurring in a subset of sCNA-bearing cells (**Figure 3.3B**). Again I have three subpopulations: the original cells (r_0), cells carrying only the sCNA (r_1) and those

carrying both (r_2). The last subpopulation has two alternatives: mutation occurring on one of the duplicated chromosome (**B_1**) or the un-duplicated chromosome (**B_2**).

C) The mutation and sCNA emerged independently, i.e., appearing in non-overlapping populations of cells (**Figure 3.3C**). This also led to three subpopulations: the original cells (r_0), cells carrying only the mutation (r_1) and those carrying only the sCNA (r_2). Note that I do not consider the fourth population that carries both the mutation and the sCNA. This outcome would require that the mutation occurred twice, once in the original cells and again in the sCNA-bearing cells. Or it requires the sCNA to occur twice. Under the Maximal Parsimonious assumption, recurrent appearance of the same mutation or the same sCNA is highly unlikely in the same tumor.

The three scenarios outlined above covered all the possible mutation-sCNA combinations for one-copy amplification without recurrence. In **Figure 3.4**, I show that heterozygous deletion and CN-LOH involve similar scenarios **A**, **B** and **C**, and each leads to a similar set of three subpopulations as described by r_0, r_1, and r_2, with $r_0 + r_1 + r_2 = 1$.

3.5.3 CCF as a function of sAGP, SAF and the underlying scenario

When the (n_a, n_b) configuration and evolutionary scenario is known, CCF can be estimated from (1) the pre-estimated sAGP of the sCNA (denoted *p* hereafter for simplicity) on which the mutation occurs, and (2) the observed allele frequency, SAF, of the somatic mutation (denoted *f* hereafter). In the following I derive the estimation procedure for heterozygous

duplication ($n_a = 2$, $n_b = 1$) and formulize general expressions for all sCNA types.

For amplification, in scenario **A_1**, n_t=3, the average total copy number $n_t = 2\times(1-p) +$ $\square_t \times p = 2 + p$. The sAGP $p = r_2$. The SAF $f = (r_1 + 2r_2)/(2 + p)$. This led to the expression $r_1 = f * (2 + p) - 2r_2$. Since $CCF = r_1 + r_2$, I have

$$CCF^{A_1}(f, n_b, n_t, p) = f * (2 + p) - r_2 = f * (2 + p) - p \quad (1)$$

In **A_2**, the situation is similar to **A_1** except that $f = (r_1 + r_2)/(2 + p)$. This led to $r_1 = f * (2 + p) - r_2$, and

$$CCF^{A_2}(f, n_b, n_t, p) = f * (2 + p) \quad (2)$$

In **B_1 and B_2**, the sAGP: $p = r_1 + r_2$. The SAF: $f = r_2/(2 + p)$. This led to $r_2 = f * (2 + p)$. Since $CCF = r_2$, I have

$$CCF^{B}(f, n_b, n_t, p) = f * (2 + p) \quad (3)$$

In **C_1 and C_2**, the sAGP: $p = r_2$. The SAF: $f = r_1/(2 + p)$. This led to $r_1 = f * (2 + p)$. Since $CCF = r_1$, I have

$$CCF^{C}(f, n_b, n_t, p) = f * (2 + p) \quad (4)$$

Note that equations (2), (3) and (4) are identical. Thus even if I do not know how to distinguish among scenarios **A_2**, **B** and **C**, CCF still has the same dependency on sAGP and SAF, and can be estimated as long as I can recognize **A_1** and **A_2/B/C**. Thus CCF identifiability is easier to achieve than scenario identifiability.

In the general copy number configuration of n_a and n_b, for scenarios A_1, A_2, B and C I have

$$\square CF^{A_1}(f, n_b, n_t, p) = n\times f - p\times n_a + p \quad (5)$$

$$CCF^{A_2}(f, n_b, n_t, p) = n_t\times f - p\times n_b + p \quad (6)$$

$$CCF^{B/C}(f, n_b, n_t, p) = n_t \times f \quad (7)$$

with $n_t = 2\times(1-p) + n_t \times p$, is the averaged copy number at the locus.

Thus, for a given pair of mutation and sCNA, with known SAF and sAGP values, once I know which scenario applies I can use Eqs. 1-7 to estimate CCF with a statistical approach as described in (Landau et al). The distribution of CCF is modeled as Binomial:

$$\Pr(CCF = x) \propto Binomi\big(S \big| N, G(x, p, \Theta)\big)$$

where S is the read count for the somatic allele and N is the total read depth. $G(\bullet)$ is expected value of SAF given CCF value x, sAGP value p, and lineage scenario Θ. G is simply obtained by reversing the CCF expressions described in the main text (Eqs. 1-7). I assume a uniform prior on x and the expectation and variance of CCF can be calculated as:

$$EXP(CCF) = \frac{\int_0^1 Binom(S|N,G)xdx}{\int_0^1 Binom(S|N,G)dx}$$

$$Var(CCF) = \frac{\int_0^1 Binom(S|N,G)x^2dx}{\int_0^1 Binom(S|N,G)dx} - EXP(CCF)^2$$

3.5.4 Joint distribution of (p, f) and scenario identifiability

By definition, f and p are both bounded by (0,1). In any tumor, however, the possible range of f is constrained by p as well as by the sCNA type and the individual scenarios. For example, in scenario **B** of amplification, the mutation occurs in a subset of sCNA-bearing cells, thus f is always less than p (in this case it is always less than 0.5 p). As I show below, the attainable

joint distributions of (*p, f*) differs among different scenarios and, importantly, this offers the possibility to infer the most likely scenario for a given sCNA-mutation pair based on their (*p, f*) values. Further, some "zones" of the (*p, f*) space overlap with multiple scenarios, thus if the observed (*p, f)* fall into these zones, it is impossible to unambiguously identify the exact evolutionary scenarios. Even then, however, because different scenarios sometimes have the same expression of CCF as a function of (*p, f*), CCF may still be uniquely estimated. In the following I derive the scenario-dependent (*p, f*) joint distributions using heterozygous amplification as example.

In $\mathbf{A_1}$, for a given *p*, the observed *f* of the mutation depends on the relative abundance of the r_0 and r_1 populations (**Figure 3.3**). When $r_0 = 0$, the mutation occurred so early that all the diploid cells carry the mutation and belong to the r_1 subpopulation. $r_1 = 1 - p$, and *f* reaches its upper limit:

$$f_h^{A_1} = \frac{1-p+2\times p}{n_t} = \frac{1+p}{2+p} \quad (8)$$

where $n_t = 2\times(1-p) + 3\times p$, is the averaged total copy number for the sCNA. On the opposite end of the situation is $r_1 = 0$, when the sCNA occurred immediately after the mutation such that none of the diploid cells carries the mutation. The lower limit of SAF is reached:

$$f_l^{A_1} = \frac{2p}{2+p} \quad (9)$$

If I plot the possible (*p, f*) combinations in an *p-f* plot with *f* on the vertical axis, under scenario $\mathbf{A_1}$, the observed f is bounded by (2p/(2+p), (1+p)/(2+p)), where $p \in (0,1)$, forming the zone marked $\mathbf{A_1}$ in **Figure 3.5A**.

For $\mathbf{A_2}$, I similarly obtain:

$$f_h^{A2} = \frac{1-p+p}{n_t} = \frac{1}{2+p} \quad (10)$$

$$f_l^{A2} = \frac{p}{2+p} \quad (11)$$

The observed f for A2 is bounded by (p/(2+p),1/(2+p)).

For **B**, f depends on the relative abundance of the r_1 and r_2 populations, and the expressions are

$$f_h^B = \frac{p}{n_t} = \frac{p}{2+p} \quad (12)$$

$$f_l^B = 0 \quad (13)$$

The f is bounded by [0, p/(2+p)].

For **C**, the upper limit of f is reached when $r_0 = 0$, $r_1 = 1 - p$, and

$$f_h^C = \frac{1-p}{n_t} = \frac{1-p}{2+p} \quad (14)$$

$$f_l^C = 0 \quad (15)$$

The f is bounded by [0, (1-p)/(2+p)].

The results for CN-LOH and deletion are shown in **Figure 3.5B-C**.

The task is to use the observed somatic allele frequency (f) and sAGP value to determine the most likely scenario among the four scenarios described in the main text. I assume that f has a uniform prior, U(0,1), and I am interested in calculating the likelihood that the sSNV occurred before the sCNA, given the copy number configuration (n_b, n_t), known sAGP (p_s), and the observed allele counts. Let f_0 denote the true f. The probability of observing S count of the somatic allele is model by Binomial(f_0, N) and the likelihood of each scenario is the probability of observing S given the scenario is true, integrated over all the possible values of

f_0:

$$p_X = L(Scenario\ X|p, n_b, n_t, N, S) = \Pr(X, p, n_b, n_t, N)$$

$$= \int \Pr(S|f_0, N) \times \Pr(f = f_0|X, p, n_b, n_t)\, \mathrm{d}f_0 = \int_{f_l^X}^{f_h^X} \Pr(S|f_0, N) df_0$$

where X is $\mathbf{A_1}$, $\mathbf{A_2}$, **B** or **C**, representing one of the four scenarios, and f^X_h, f^X_l are computed according to Eqs. 8-15. I then compute the summation of p_X:

$$P = p_{A1} + p_{A2} + p_B + p_C$$

and normalize each likelihood using P:

$$\tilde{p}_X = \frac{p_X}{P}$$

I calculate the normalized probability for each scenario, as well as all the possible combinations of multiple scenarios. For example, the probability of either scenario A1 or C is $\tilde{p}_{A_1C} = \tilde{p}_{A_1} + \tilde{p}_C$. There are in total 2^4 - 1=15 possible combinations. If the normalized probability of any of the four scenarios is greater than 0.95, the SNV is assigned to the corresponding scenario. If none of the single-scenario probability exceeds 0.95, I ask if any of the six two-scenario combinations have probability > 0.95. If this step fails, I next examine the four possible three-scenario combinations, and so forth. If all the above steps fail, I report the SNV scenario $\mathbf{A_1/A_2/B/C}$, and no unique CCF can be estimated in this case.

To state the full estimation procedure: when (f, n_b, n_t, p) are known for a mutation-sCNA pair, if the (*p, f*) combination identifies a unique scenario according to **Figure 3.5**, CCF is calculated using Eqs. 3-7. If the (*p, f*) combination overlaps with multiple scenarios, CCF may still be calculated if the expressions are the same across the undistinguishable scenarios. Lastly, when the CCF expressions are different among the applicable scenarios, CCF cannot

be uniquely determined, however its 2 or more possible values can still be obtained as valid alternatives. In implementation, as SAF is a random variable with confidence level depending on read depth, there is always uncertainty as to which scenario the observed (*p, f*) belongs.

3.6 Validation and performance

3.6.1 Performance of sAGP inference

I first tested the performance of *CHAT* in sAGP estimation. I simulated LRR and BAF values for a series of sCNA datasets with two aneuploid tumor populations, which are mixed with the euploid population. The first population is the dominant clone, with an assigned sAGP value of $p_{dom} \sim [0.1,0.2,\ldots,1.0]$. The second population is a minor clone, with an assigned sAGP value of $p_{sub} \sim [0,0.1,\ldots p_{dom}\text{-}0.1]$. The fraction of the euploid population is $1- p_{dom} - p_{sub}$. In all, there are 55 $p_{dom} - p_{sub}$ combinations; and for each, I simulated 200 euploid segments ($n_b = 1, n_t = 2$, sAGP = 0) and 200 sCNA segments, of which 133 (about 2/3) were assigned to the dominant clone (sAGP = p_{dom}), and the remaining 67 were assigned to the minor clone (sAGP = p_{sub}). Within each clone, the sCNAs were assigned to one of four copy number configurations with the following ratios: 2/7 for deletion ($n_b=0,n_t=1$), 2/7 for CN-LOH ($n_b=0,n_t=2$), 2/7 for amplification ($n_b=1,n_t=3$) and 1/7 for balanced doubling ($n_b=2,n_t=4$). The BAF and LRR values were generated using the assigned sAGP and copy

number configuration with the following formula:

$$BAF = \left|0.5 - \frac{p \times n_b + 1 - p}{n_t}\right| + Normal(0, \sigma_{BAF}) \quad (16)$$

$$LRR = log_2 n_t - 1 + Normal(0, \sigma_{LRR}) \quad (17)$$

where p stands for sAGP, and n_t is the averaged total copy number for the local segment: $2(1-p) + n_t \times p$. σ_{BAF} and σ_{LRR} are the standard deviation values of the per-segment BAF and LRR, respectively. For the Affymetrix 6.0 platform, the per-SNP standard deviation for BAF is about 0.05, and for LRR is about 0.25 (our observation). Thus the choice of σ_{BAF}=0.01 and σ_{LRR}=0.04 is equivalent to an sCNA of approximately 36 SNP markers. For a 1 million SNP platform, 36 SNPs cover approximately 110 kb, therefore ours are conservative choices for sCNAs 110 kb or longer, profiled by 1 million SNPs or more.

After generating the BAF and LRR values using Eqs. 16-17 for the 400 segments for each of the 55 p_{dom} - p_{sub} combinations, I applied *CHAT* to estimate sAGP, n_b, and n_t for each simulated segment, and evaluated performance by reporting (1) percent of cases of mistaken estimation of sCNA configuration (error in either n_b or n_t) (**Figure 3.6A, top row**) for dominant and minor clonal events, and (2) the median absolute deviation of the estimated sAGPs from the assigned p_{dom} or p_{sub} for the dominant and minor clones, for either the segments with correct (nt, nb) identification (**Figure 3.6A, middle row**), or all segments (**Figure 3.6A, bottom row**). With all of these performance metrics, the errors are the largest when with the clonal or subclonal sAGPs are small. The overall errors are small in most situations, suggesting that *CHAT* worked well in recovering the sAGP, n_b, and n_t values.

3.6.2 Performance of CCF prediction

Of the 55 p_{dom} - p_{sum} combinations described above I selectively tested CCF inference in four cases: p_{dom} - p_{sum} ~ (0.9,0.8), (0.9,0.4), (0.5,0.3), and (0.3,0.1). For each case, I simulated 4,000 SNVs, of which ~2,000 fall in the 200 euploid segments, and the other 2,000 fall in the 200 sCNA regions, with sAGP-n_b-n_t assignment as described above. In effect I assume that the euploid intervals account for 50% of the genome. For all downstream inferences, I used the sAGP, n_b and n_t estimated by *CHAT*. If the SNV falls in a euploid region, the assigned SAF was randomly drawn from uniform(0,0.5) and the assigned CCF = SAF*2. If it falls in an aneuploid region, I randomly choose the lineage scenario from (**A_1**, **A_2**, **B**, **C**) according to the local copy number a configuration. If the sCNA is a CN-LOH or balanced doubling region, I limit the scenarios to (**A_1**, **B**, **C**). The upper and lower limits of the chosen scenario were determined using Eqs.8-15. SAF values were then randomly drawn from uniform(f_l, f_h), where f_l and f_h were the lower and upper limits. "Known" CCF values were computed using Eqs. 1-6 in the main text. Lastly, I simulated the allele counts in two steps. For a mean read depth k, the actual coverage at a given SNV, N, was sampled from N ~ Poisson(*k*). When N and f were assigned, the count of the somatic mutation allele was sampled from Binomial(*f*, N). Based on the estimated sAGP, copy number configuration and the simulated somatic allele counts I used *CHAT* to estimate CCF. The estimated values were compared with the "known" CCF in **Figure 3.6B** for both k = 50 and k = 100.

In all eight cases (four p_{dom} – p_{sub} combinations and two k values) the Spearman's rank

correlation coefficient between the known and estimated CCF values ranged in 0.946- 0.97, indicating that *CHAT* makes accuracy CCF inference. To compare performance among SNVs in different sub-categories, I separated those falling in euploid regions from those in sCNAs, and for the latter, separated those in the major and minor subclone events, and those in different copy number status. As shown in **Figure 3.6C**, the error rates are similar across these sub-categories, not affected by dominant/minor clonal events or different sCNA types.

3.6.3 Computational requirements

I estimated the time and memory requirement of *CHAT* using the TCGA dataset for breast tumors. The time estimate below is based on allele-specific copy number data with 850K SNPs for tumor-normal pairs and whole-exome sequencing data with ~30X average coverage. For binned segmentation (~500 heterozygous SNPs per bin), it takes 2 minutes to complete the sAGP and CCF estimation for one tumor/normal pair, and it requires about 10 MB memory. For detected sCNAs, the computational time increases to an average of 12 minutes per sample pair. The above estimation is based on running R scripts with a single processor (AMD Opteron 6136, 2.4GHz with 4G RAM) and counting input file reading time. In *CHAT*, the user can apply the R package *parallel* to enable multi-thread processing. This allows the use of as many processors as available. On our server (32 AMD Opteron 6136 CPUs and 128G RAM), our test run used 14 processors on average, and it took 10 hrs (140 CPU-hours) to complete the CBS segmentation, sAGP estimation for 732 breast tumor-normal samples

and CCF estimation for 445 samples with downloaded VCF files.

3.7 Application to human breast cancer

I applied *CHAT* to estimate sAGP for sCNAs identified using Affymetrix 6.0 single nucleotide polymorphism (SNP) array data for tumor and germline DNA samples from 732 breast cancer patients (The Cancer Genome Atlas Research Network, 2012b). Of these, 445 also have whole-exome sequencing data available, and I estimated CCF for SNVs.

3.7.1 sAGP distribution

I detected sCNAs using circular binary segmentation (Olshen et al., 2004) of LRR and BAF data (Li et al., 2012), resulting in the identification of an average of 261 sCNAs per tumor (range: 1 - 3,537). The median size of all sCNAs is 1.7 Mb (range 2.5 Kb – 245 Mb). On average, each tumor carries 125 sCNAs larger than 5 Mb, a size corresponding to ~1,500 SNP markers in the 850K SNP array. Given this sCNA size range, I re-calculated sAGP for genomic bins containing 500 heterozygous SNPs in the germline DNA, a bin size that is approximately 5 Mb, resulting in 502 bins per sample (range: 404 – 794) and constructed the sAGP histogram for every tumor. 87 tumors (12%) had sCNAs for <50 bins, too few for analyzing the sAGP distribution patterns. For the remaining 645 tumors I fit the sCNAs

distribution to either a uni-modal + uniform distribution or a multi-modal distribution using methods described above. In the example in **Figure 3.2C**, a three-mode distribution provides the best fit, with sAGP peaks around 0.5, 0.4, and 0.2. The highest peak corresponds to the black-colored sCNAs in **Figure 3.2A-B**, while the second and third peaks correspond to the red and green-colored sCNAs, respectively. In total I observed 392 samples (61%) with best fit by the multi-modal distribution, while 253 (39%) follow the uni-model + uniform distribution. This shows that a majority of the breast tumors analyzed by TCGA contain more than one recognizable aneuploid population, suggesting that the co-existence of more than one subclone is very common.

3.7.2 sAGP-CCF joint distribution for known cancer genes

The 445 tumors with both SNP array and sequencing data have an average of 311 somatic mutations per tumor with CCF values (range: from 15 to 4235, after counting the 8.8% loss due to sCNAs with un-estimable sAGP). While 48% of these mutations fall into a zone with overlapping scenarios, 93% of them have a unique mathematical expression and can produce a valid CCF estimate (**Figure 3.7**). The remaining 7% are assigned missing CCF values due to scenarios with conflicting CCF estimates.

The calculation of sAGP for most sCNAs and CCF for most SNVs makes it possible to examine the joint distribution of clonality for these two types of genome aberrations. A "CCF vs. sAGP" plot can be created for all copy number and mutation events in a single tumor, or

for events affecting a single gene of interest across many tumors. For a given gene, if a sample does not have any somatic mutation in the gene, I assign CCF = 0. Likewise if the copy number of the gene is normal, I assign sAGP = 0. **Figure 3.8A** shows a heatmap depicting the density of CCF and sAGP joint distribution for all events in a hypothetical sample (or for a hypothetical gene across all samples). In this two-dimension space, the "hot" peak near the origin (0,0) is typical for most genes, affected by neither somatic mutation nor sCNA. The peak in the upper left (near the sAGP-axis) contains genes with highly clonal CNAs but carrying either no mutation or mutations of low clonality. A plausible interpretation is that for some of these genes, sCNA is a possible driver event. Similarly, the peak at the lower left (near the CCF-axis) contains genes with highly clonal somatic mutations and low-clonality sCNAs. Lastly, genes in the upper-right peak have both high sAGP and high CCF values, suggesting that both copy number changes and somatic mutations may have occurred at very early stages of tumor development, and their joint appearance may be necessary to act as a driver event.

Figure 3.8B allows close inspection of relative clonality between sCNA and mutations for four genes known to be related to breast cancer (The Cancer Genome Atlas Research Network, 2012b): *TP53*, *PIK3CA* and *GATA3*, which occurred in > 10% of analyzed breast tumors, and *MAP3K1*, which had mutations enriched in the luminal A subtype. For *TP53*, there are two noticeable high-density "zones" in the heatmap: one along the sAGP-axis, the other at the upper right, indicating two groups of tumor samples: *TP53* CNA-only and *TP53* CNA/mutation, respectively. This pattern, when stratified by the four PAM50 subtypes (The

Cancer Genome Atlas Research Network, 2012b, Parker et al., 2009) (**Figure 3.8C)**, shows that the *TP53* CNA/mutation group is enriched in the Basal and HER2 subtypes (accounting for 72 of 94 Basal and HER2 tumors), whereas the *TP53* CNA-only group is enriched in the Luminal-A (94 of 105), and to a lesser degree, the Luminal-B subtypes (44 of 67). In comparison, the other three genes have not only the CNA-only and CNA/mutation groups, but also a third, mutation-only group near the CCF-axis. **Figure 3.5D** shows that for *PIK3CA*, the mutation-only group occurs almost exclusively in the Luminal-A and –B subtypes.

The CCF - sAGP plot can also be used to compare the clonality distribution between a pair of genes. In **Figure 3.9**, *TP53* and *PIK3CA* are shown in red and blue symbols, respectively, with the lines linking the two genes for the same samples. There are three notable patterns of *TP53* - *PIK3CA* clonality. First, samples marked by the black lines have both sCNA and mutation in *TP53* but no aberration in *PIK3CA*. Second, samples marked by the red and green lines tend to have sCNA for both *TP53* and *PIK3CA* and at comparable sAGP, but only mutation in *TP53* (red lines) or *PIK3CA* (green). Third, samples marked by the blue lines had high clonality for *TP53* CNA, but not its mutation, and high clonality for *PIK3CA* mutation, but not its CNA, suggesting co-occurrence of aberrations of these two genes but involving different variant types. These patterns are subtype-specific: Pattern 1 is enriched in the Basal subtype (OR=4.6 compared to the other three subtypes, P=0.0001 by Fisher's exact test, for red; OR=1.2, P=0.67, for green), so is Pattern 2 (OR=5.3, P=6.4e-8,). Most remarkably, Pattern 3 is almost exclusively found in the Luminal A subtype (OR=56, P=4.4e-9).

3.8 Improvements of *CHAT* comparing with previous methods

While *CHAT* does not solve all the issues facing the cancer genome deconvolution problem, it attempts to overcome several important compromises or simplifying assumptions that underlie other methods. First, *oncoSNP* (Yau et al., 2010) and *ThetA* (Oesper et al., 2013) are designed to estimate sCNA clonality, but they do not address the clonality of somatic mutations. Second, Ding et al. (Ding et al., 2012) used kernel density estimation method to characterize somatic mutations, but only focused on those in the euploid regions of genome, staying clear of the complicated relationship between SNV and sCNAs. Third, *ABSOLUTE* infers clonality for both sCNA and mutations but only designate subclonal events, stopping short of quantitative estimation. This method was extended in Landau et al. (Landau et al., 2013) to estimate CCF for somatic mutations even if they are subclonal, but the algorithm only considers the case where sCNA occurred before SNV, equivalent to our **scenario B (Figure 3.3 and 4)**, and further assumes that the copy number was altered by only one in the sCNA. In this regard, *CHAT* considers a wider array of possible scenarios. Fourth, *EXPANDS* (Andor et al., 2014)works with next-generation sequencing data and jointly estimates the absolute DNA copy number, clonality of somatic mutations, and that of sCNAs. However, this method only considers **scenario $\mathbf{A_1}$** and without the intermediate r_1 population. In effect, it assumes that the mutation and sCNA occur at the same instance and are in phase. Fifth, *PyClone* (Roth et al., 2014) infers clonality of somatic mutations and performs phylogenetic analysis. It receives as input the integer copy number profiles estimated from other methods,

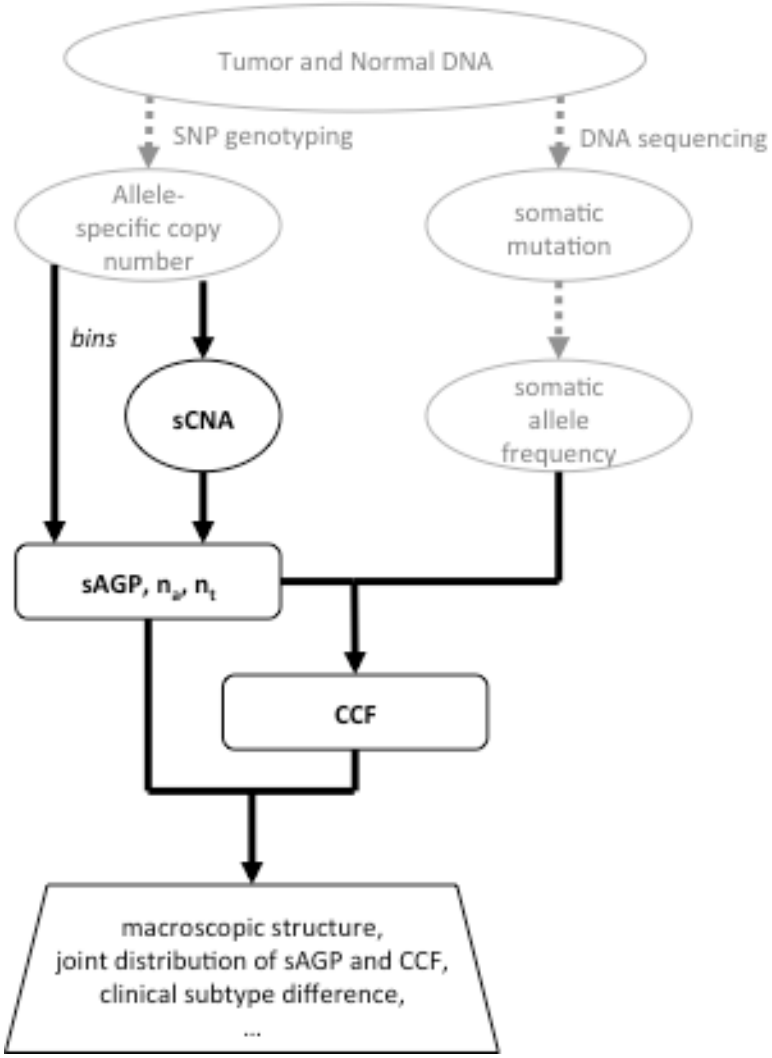
but only considers **scenarios A and B**, disregarding the possibility of a branching lineage. Furthermore, for scenario **A**, it assumes co-occurrence of SNV and sCNA, thus also ignoring the r_1 population. In short, the first key contribution of *CHAT* is in providing a general mathematical framework that enumerates the complete set of scenarios covering the possible order and phase of the sCNA and the single-base changes. Like many of the previous methods, *CHAT* has its own limitations, primarily in being unable to resolve extremely complex events such as three-way mixing or above. It models two-population mixing at each genomic region (a gene, an sCNA, or a bin) and works best when the tumor has not experienced extensive and repeated copy number alterationa. In the TCGA breast tumor dataset I found that 9.3% of sCNAs do not follow the regional two-way mixing model and do not allow sAGP estimation. For the other, permissible sCNAs, *CHAT* can proceed, and is able to infer the coexistence of two or more subpopulations by analyzing the distribution of sAGP or CCF values. I wish to re-emphasize that while *CHAT* invokes two-way mixing for each individual genomic region, it is not limited to infer the presence of only two populations of cells. Globally, the number of peaks in the sAGP or CCF distribution has no restriction, and can be very high when the signal-to-noise ratio is improved, such as with ultra-deep sequencing data (e.g., Shah et al., 2012).

3.9 Summary

In this work, I developed a computational framework to estimate clonality for both sCNAs

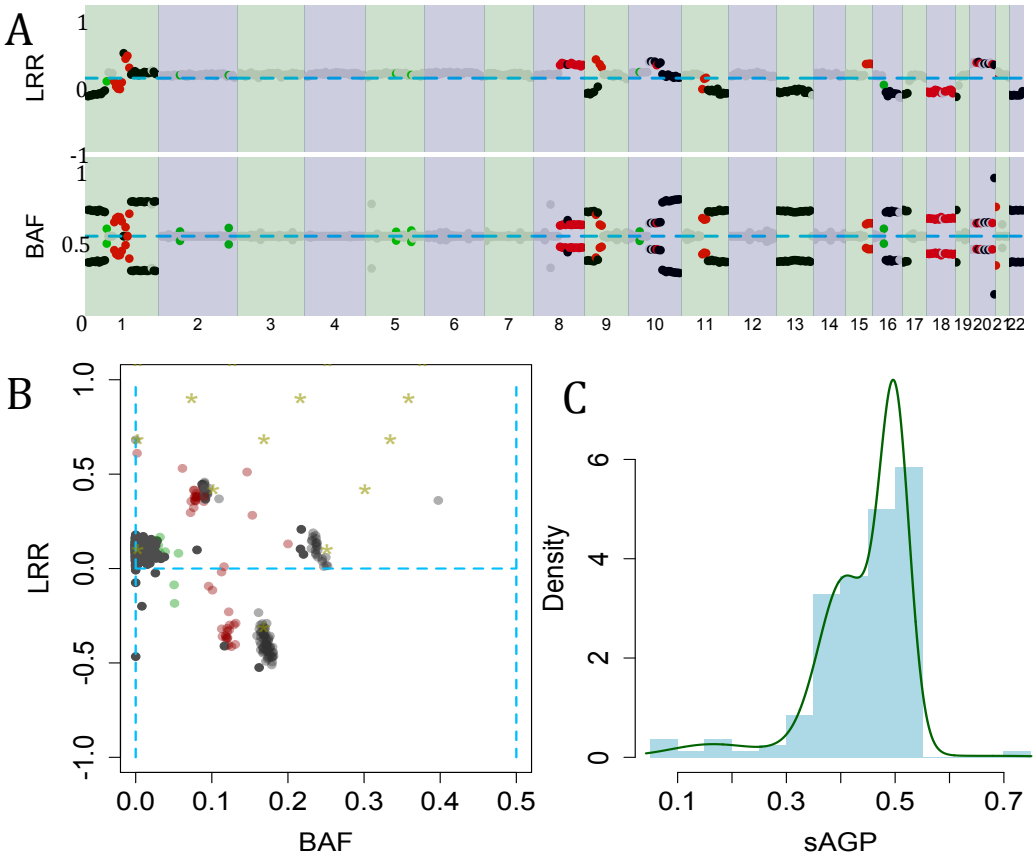
and SNVs. It is built on previous methods both by us (Li et al., 2012) and by others. It lifted several unrealistic assumptions in previous methods and clarified some ambiguous concepts. A second contribution of *CHAT* is in systematically assessing the input data combinations that lead to "unidentifiable zones", in which the CCF, or "scenarios" (i.e., the evolutionary order and phase of the sCNA and SNV), cannot be resolved even with perfect data. I found that in many situations, even if the evolution scenario is undetermined, CCF values can still be estimated. The ability to objectively evaluate the power of inference in any given dataset is an important part of method development.

Figure 3.1: Schematic pipeline of tumor subclonality using CHAT.



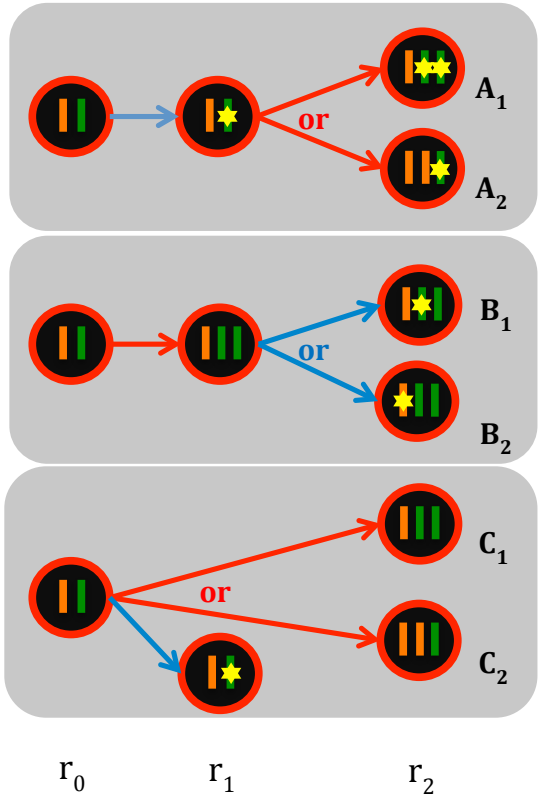
DNA extracted from tumor mass and paired normal sample were sent for both allele-specific copy number profiling and DNA sequencing. CHAT provides two alternative ways to define the spatial units for sAGP analysis: by natural DNA segmentation (CBS) or fixed number of heterozygous BAF marker bin. Inference on CCF and lineage scenarios relies on two sources of input: sAGP estimations with absolute copy number configuration, and the allele counts from the sequencing data for each somatic mutations. A wide spectrum of downstream analysis are available with the rich information inferred by CHAT, including the macroscopic subclonal structures of tumor populations, joint analyses of sAGP and CCF for each cancer gene or for each tumor sample, interactions between sAGP and CCF values for different tumor subtypes, etc.

Figure 3.2: Evolution model inference for primary tumor sample TCGA-A1-A0SD.



A: Scatter plot for binned segments of BAF and LRR showing different levels of contraction even within same type of sCNA. **B**: BAF-LRR plot for the same sample. **C**: MCMC fitting of sAGP distribution reveals three distinct modes, peaking around 0.5,0.4 and 0.2. The segments in each peak are colored black (0.5), red (0.4) and green (0.2) in A and B. Segments sharing similar sAGP values and clustered within same peak are likely carried by a same group of cells, namely subclone. In **C**, sAGP distribution is indicated by light blue histogram, while the DP fitted density is shown in dark green line.

Figure 3.3: Paradigms for lineage scenarios A to C for heterozygous amplification.



In scenario$\mathbf{A_1}$, three population of cells are modeled: euploid cell without mutation, euploid cell with mutation (hexagon star) and aneuploid cell with mutation. When mutation occurred before amplification of the green allele, both alleles carry that mutation. r_0,r_1 and r_2 are the fraction of each of the corresponding cell population (same is true to other scenarios) and sum up to 1. $\mathbf{A_2}$ is similar to $\mathbf{A_1}$, except that the mutation occurred on the unamplified allele (orange). For scenario **B**, there are two equally possible cases: mutation occurred on the amplified ($\mathbf{B_1}$) or unamplified ($\mathbf{B_2}$) allele. The formulas to compute CCF for either case are identical. For scenario **C**, where mutation and sCNA independently occurred on difference lineages, there are also two possible cases: mutation occurred on maternal($\mathbf{C_1}$) or paternal($\mathbf{C_2}$) allele in the euploid cells, and the formulas for either case are also identical. Blue arrow: mutation occurrence; red arrow: sCNA occurrence.

Figure 3.4: Lineage scenarios for CN-LOH (A) and heterozygous deletion (B).

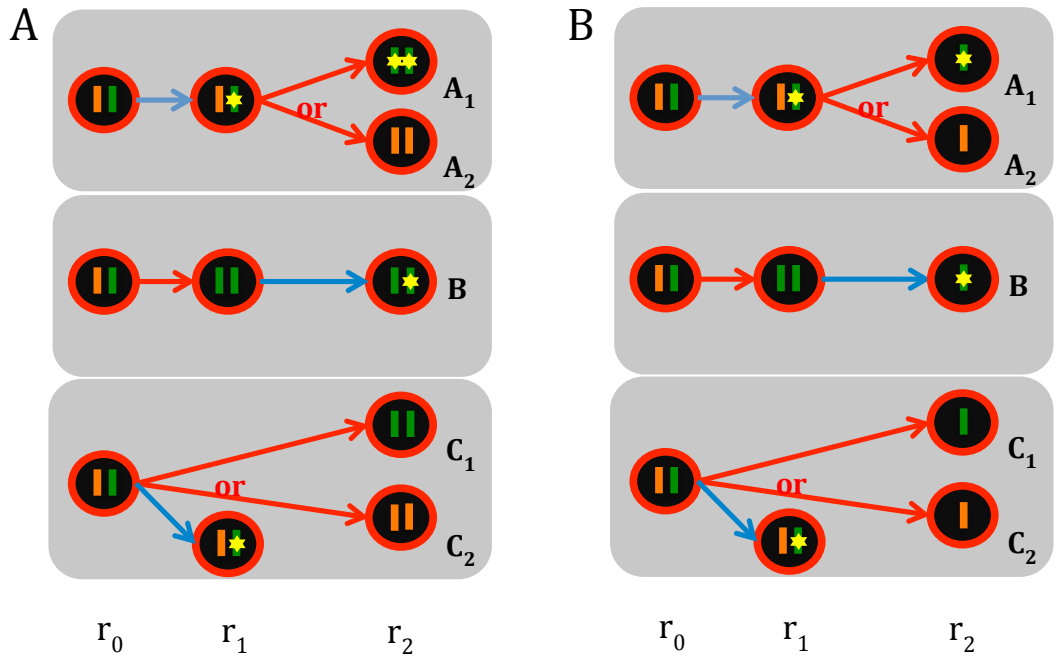
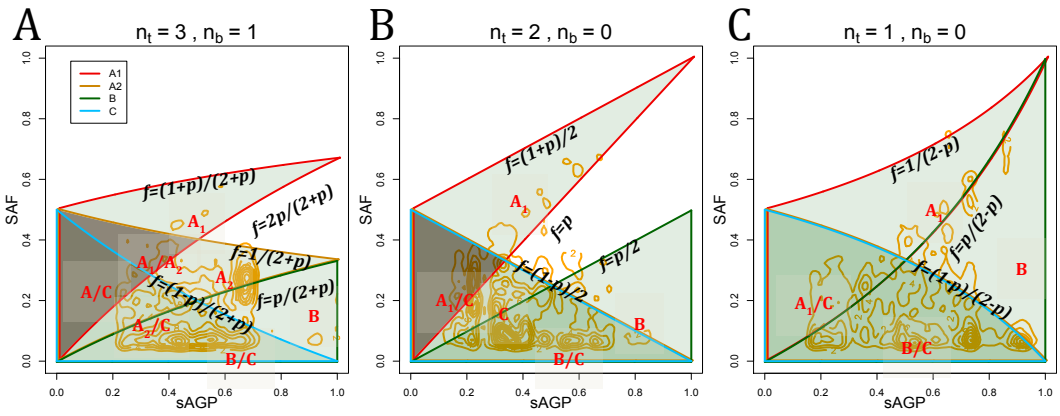
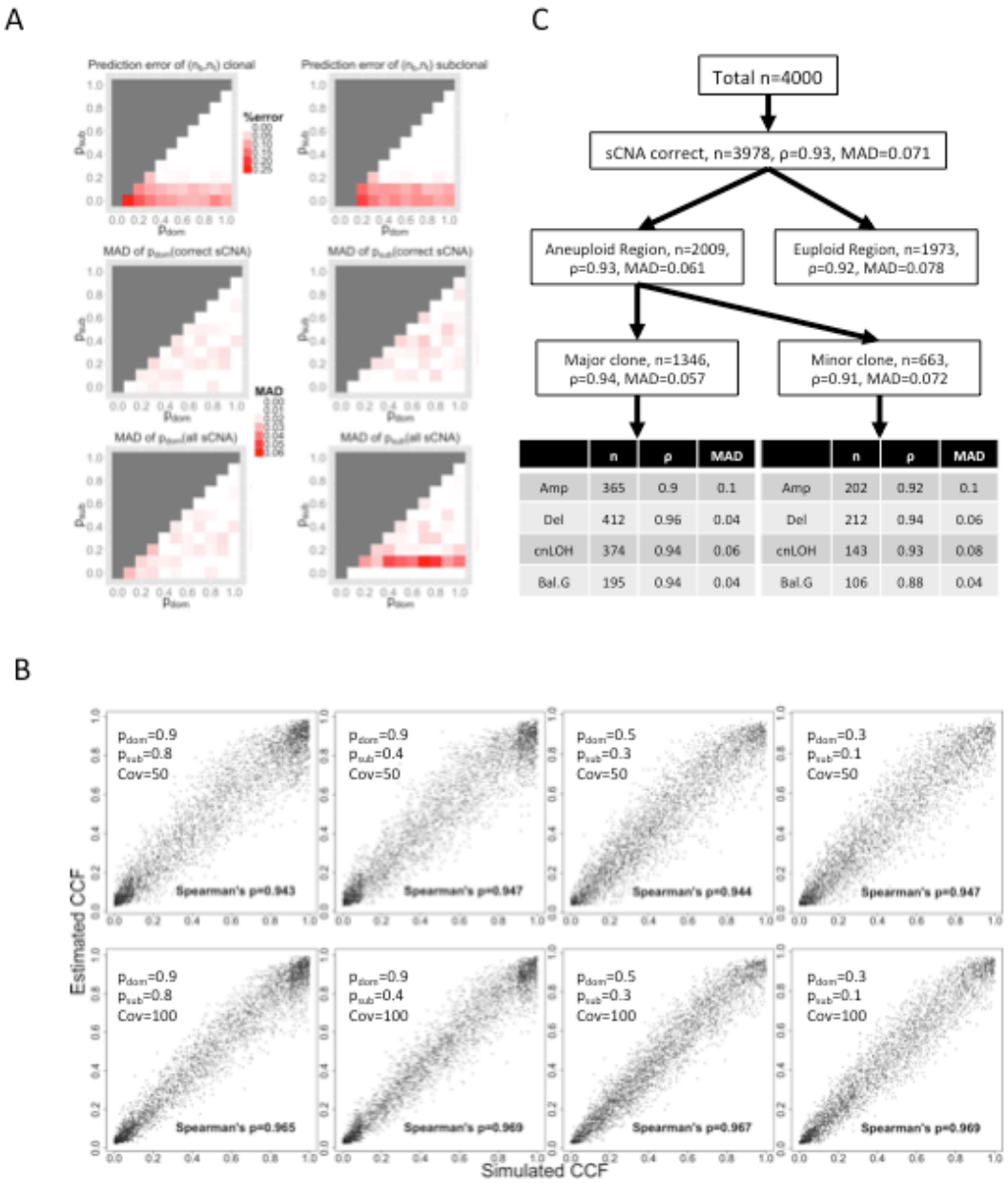


Figure 3.5: Identifiable zones



Identifiable zones for CCF estimation in case of hemizygous amplification (**A**), cn-LOH (**B**) and hemizygous deletion (**C**), for four temporal scenarios described in the main text.Somatic mutations from 201 diploid tumor samples with percent on point (PoP) greater than 0.05 (a measure of prediction accuracy from (Li et al., 2012)) were selected for the analyses. Lineage scenarios are bordered with different colors as displayed in the legend. Variants with coverage lower than 20 or SAF smaller than 0.05 were discarded. 3382 mutations from hemizygous amplification, 2008 from cn-LOH and 4662 from hemizygous deletion regions were plotted respectively for each sCNA type against the corresponding sAGP values of the DNA segments. Theoretical boundaries of SAF values in each scenario from Eq(8)-(15) were overlaid on the plot to display the unique and unidentifiable zones. Each lineage scenario is bordered by a different color, as indicated in the legend of **A**. Text boxes with black color labels the regions on the plot their attributes. Single letter in the box indicates that the region is uniquely assigned to that scenario, while multiple letters indicate the region is indistinguishable between the corresponding scenarios. Regions without text box of capital letters are theoretically impossible, and our results show that the distribution of real data agrees well with most of these regions. CCF of somatic mutations in light green areas can be uniquely estimated, while not in gray zones due to unidentifiable issue.

Figure 3.6: *In silico* validation of *CHAT* performances.



A. Performance of sAGP inferences. Upper row: percent of error in estimated nb or nt, for the dominant (left) and subclonal sCNAs (right), as described in Materials and Methods, Sec8on 7a. Middle row: the median absolute difference (MAD) between predicted and simulated sAGP values for sCNAs with correctly identified (n_b, n_t), or for all sCNAs (Bottom row). The p_{sub}=0 row of the lower-right and middle-right panels had zero error because when p_{sub}=0 there is only one clone in the tumor population and all subclonal sCNA segments have correctly estimated sAGP = 0. B. Performance of CCF inference. Shown are scatter plot of simulated and estimated CCF for four p_{dom} – p_{sub} cases and two coverage values: Cov=50 (upper panels) and 100 (lower panels). C. Comparison of CCF inference accuracy among different SNV categories: euploid vs. aneuploidy regions; and in the

latter, between the dominant and the minor clones. Lastly, SNVs were divided by sCNA types. The tested case has the following parameter settings: p_{dom}=0.9, p_{sub}=0.6, coverage=50, number of SNV sampled=4,000, number of sCNA sampled=200.

Figure 3.7: Distribution of the percentage of somatic mutations associated with a unique scenario (black) and the additional percentage with unique CCF estimates (red). From left to right are the results for 445 breast tumor samples, ordered by the unique-scenario percentage.

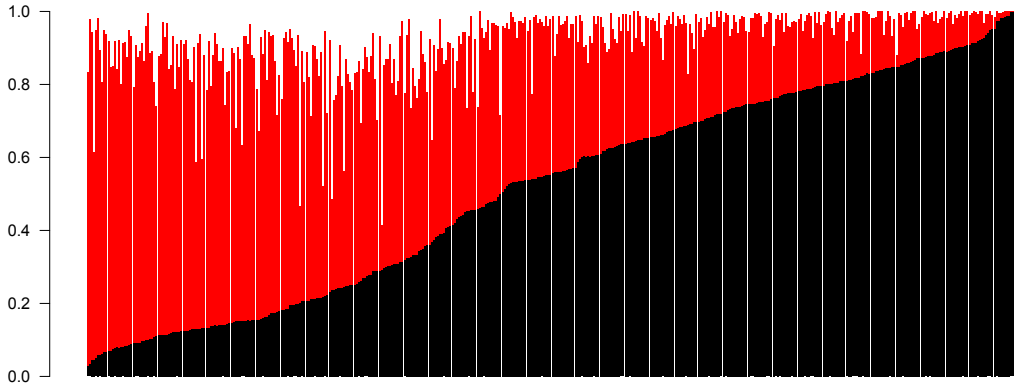
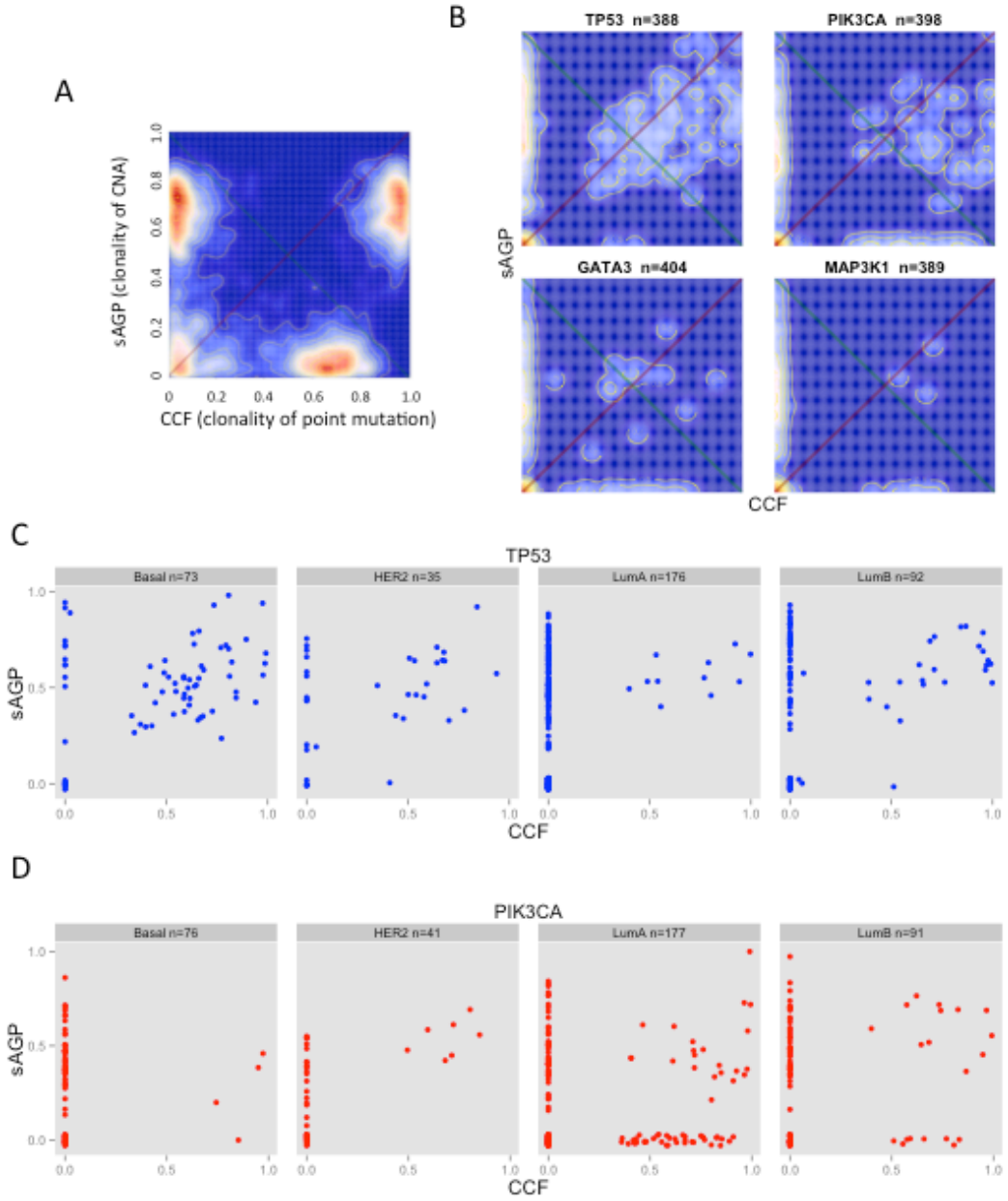
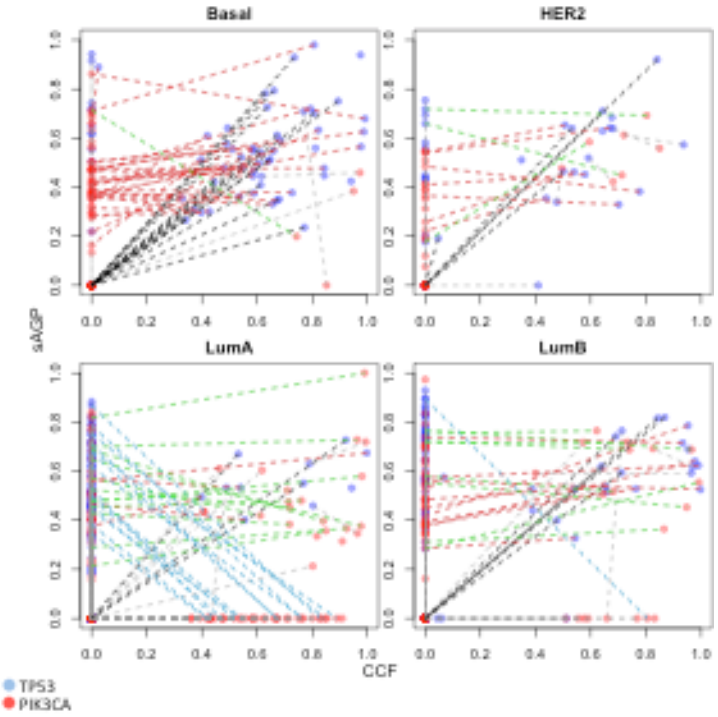


Figure 3.8: Single gene summary for sAGP-CCF joint distribution for 445 BRCA samples.



A: A made-up example showing characteristic density peaks on the heatmap. **B**: Realization of **A**) for four breast cancer related genes: TP53, PIK3CA, GATA3 and MAP3K1. **C-D**: Scatter plot of sAGP-CCF for two genes, TP53 and PIK3CA, stratified by PM50 BRCA subtypes. sAGP values for euploid regions are added a small random noise for visualization purpose. Numbers after gene symbols in **B** are the number of samples with both sAGP and CCF estimable for the gene across 445 tumors. For each subtype, in **C** and **D**, the number indicates the same. To note, I excluded 7 Normal-like samples due to low count.

Figure 3.9: Two-gene CCF-sAGP comparison for TP53 and PIK3CA across 445 samples and stratified by PM50 gene expression subtypes.



Interactions between CCF and sAGP for TP53 are characterized by interacting types: 1. correlated CCF and sAGP value of TP53, mostly enriched in Basal subtype (dashed black line); 2. correlated sAGP values for TP53 and PIK3CA (dashed red line); 3. correlated TP53 sAGP and PIK3CA CCF, enriched in LumA subtype (dashed blue line); 4. correlated TP53 sAGP and PIK3CA sAGP, with PIK3CA somatic mutations (dark green line).

3.10 Bibliography

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A. L.*, et al.* 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-21.

Andor, N., Harness, J. V., Muller, S., Mewes, H. W. & Petritsch, C. 2014. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics,* 30**,** 50-60.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., Mclaren, S., Lin, M. L.*, et al.* 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature,* 467**,** 1109-13.

Carter, S. L., Cibulskis, K., Helman, E., Mckenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A.*, et al.* 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol,* 30**,** 413-21.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y.*, et al.* 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* 486**,** 346-52.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., Mclellan, M. D.*, et al.* 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature,* 481**,** 506-10.

Durinck, S., Ho, C., Wang, N. J., Liao, W., Jakkula, L. R., Collisson, E. A., Pons, J., Chan, S. W., Lam, E. T., Chu, C.*, et al.* 2011. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov,* 1**,** 137-43.

Escobar, M. D. & West, M. 1995. Bayesian Density-Estimation and Inference Using Mixtures. *Journal of the American Statistical Association,* 90**,** 577-588.

Fearon, E. R. & Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-67.

Fidler, I. J. 1978. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer Res,* 38**,** 2651-60.

Garraway, L. A. & Lander, E. S. 2013. Lessons from the cancer genome. *Cell,* 153**,** 17-37.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P.*, et al.* 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.

Greaves, M. & Maley, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D.*, et al.* 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell,* 148**,** 873-85.

Jara, A., Hanson, T. E., Quintana, F. A., Muller, P. & Rosner, G. L. 2011. DPpackage: Bayesian Non- and Semi-parametric Modelling in R. *J Stat Softw,* 40**,** 1-30.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A.*, et al.* 2008. Core signaling pathways in human pancreatic cancers

revealed by global genomic analyses. *Science,* 321**,** 1801-6.

Keats, J. J., Chesi, M., Egan, J. B., Garbitt, V. M., Palmer, S. E., Braggio, E., Van Wier, S., Blackburn, P. R., Baker, A. S., Dispenzieri, A.*, et al.* 2012. Clonal competition with alternating dominance in multiple myeloma. *Blood,* 120**,** 1067-76.

Knudson, A. G., Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

Landau, D. A., Carter, S. L., Stojanov, P., Mckenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L.*, et al.* 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell,* 152**,** 714-26.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A.*, et al.* 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-8.

Ley, T. J., Ding, L., Walter, M. J., Mclellan, M. D., Lamprecht, T., Larson, D. E., Kandoth, C., Payton, J. E., Baty, J., Welch, J.*, et al.* 2010. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med,* 363**,** 2424-33.

Li, B., Senbabaoglu, Y., Peng, W., Yang, M. L., Xu, J. & Li, J. Z. 2012. Genomic estimates of aneuploid content in glioblastoma multiforme and improved classification. *Clin Cancer Res,* 18**,** 5595-605.

Mcfadden, D. G., Papagiannakopoulos, T., Taylor-Weiner, A., Stewart, C., Carter, S. L., Cibulskis, K., Bhutkar, A., Mckenna, A., Dooley, A., Vernon, A.*, et al.* 2014. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell,* 156**,** 1298-311.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D.*, et al.* 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* 2012. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

Nowell, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

Oesper, L., Mahmoody, A. & Raphael, B. J. 2013. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol,* 14**,** R80.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics,* 5**,** 557-72.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z.*, et al.* 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol,* 27**,** 1160-7.

Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E. & Stern, M. H. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol,* 10**,** R128.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Cote, A. & Shah, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods,* 11**,** 396-8.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K.,

Haffari, G., *et al.* 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature,* 486**,** 395-9.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature,* 498**,** 236-40.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., *et al.* 2006. The consensus coding sequences of human breast and colorectal cancers. *Science,* 314**,** 268-74.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C. & Tavare, S. 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A,* 110**,** 4009-14.

The Cancer Genome Atlas Research Network 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455**,** 1061-8.

The Cancer Genome Atlas Research Network 2011. Integrated genomic analyses of ovarian carcinoma. *Nature,* 474**,** 609-15.

The Cancer Genome Atlas Research Network 2012a. Comprehensive genomic characterization of squamous cell lung cancers. *Nature,* 489**,** 519-25.

The Cancer Genome Atlas Research Network 2012b. Comprehensive molecular portraits of human breast tumours. *Nature,* 490**,** 61-70.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., *et al.* 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., *et al.* 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell,* 17**,** 98-110.

Vogelstein, B. & Kinzler, K. W. 1993. The multistep nature of cancer. *Trends Genet,* 9**,** 138-41.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., *et al.* 2007. The genomic landscapes of human breast and colorectal cancers. *Science,* 318**,** 1108-13.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., *et al.* 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell,* 148**,** 886-95.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., *et al.* 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467**,** 1114-7.

Yates, L. R. & Campbell, P. J. 2012. Evolution of the cancer genome. *Nat Rev Genet,* 13**,** 795-806.

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O. & Holmes, C. C. 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol,* 11**,** R92.

Chapter 4. *STRfinder*: A general tool for detecting and genotyping short tandem repeat variation using paired-end next-generation sequencing data

4.1 Introduction

Short tandem repeat (STR) in the genome refers to the consecutive occurrence of the same 2-6 DNA base pairs many times in a row. Its allelic variation represents an important class of genetic variation in many genome systems, including the human genome and the genomes of many cancers. STR variants can affect protein structure or gene regulation (Gemayel et al., 2010, Kozlowski et al., 2010, Bolton et al., 2013, Sawaya et al., 2013), and have been implicated in several inherited diseases in humans (Mirkin, 2007). Meanwhile, cancer researchers have realized since decades ago that the instability in STR alleles may be associated with certain cancers (Wooster et al., 1994), such as colorectal cancer (Markowitz et al., 1995, Parsons et al., 1995, Popat et al., 2005). Further, CAG repeat polymorphism in androgen receptor (AR) is implicated prostate cancer (Nelson and Witte, 2002) and male breast cancer (MacLean et al., 2004). Highly polymorphic STR loci have also been used as DNA fingerprinting markers for forensic identification purposes (Keats et al., 2012). In medical genetics, STRs in coding regions sometimes undergo abnormal expansions and the alleles with high numbers of repeat units can increase the risk of certain diseases. There have been many well

documented developmental or neuro-degenerative disorders that are caused by STR expansions (Gatchel and Zoghbi, 2005), including the Huntington's disease (Walker, 2007), Fragile-X syndrome (Pearson et al., 2005), Machado-Joseph disease (MJD) (Paulson, 2012), and some types of ataxia (Orr et al., 1993, Pulst et al., 1996, Campuzano et al., 1996).

Several methods have been developed to annotate STR location/length (Benson, 1999, Smit, 1996-2004) in well-assembled sequences, such as the human reference genome. However, a reference genome represents a consensus sequence, and does not capture STR variations in a population. In this regard, we still lack robust methods for detecting STR variation and genotyping individual samples. Experimentally, the detection of STR remains error-prone and difficult to scale up. Traditional Sanger sequencing still serves as a gold standard, but it cannot be efficiently applied to genome-wide scans or in larger sample cohorts. In recently years, with the arrival of the next-generation sequencing (NGS) technologies, it becomes feasible to collect DNA sequence data over a large sample and sometimes in exome-wide or genome-wide fashion. Specifically for the short-read NGS data, two methods, lobSTR (Gymrek et al., 2012) and RepeatSeq (Highnam et al., 2013), have been recently developed to detect STR variation. Both methods rely on an existing database of known STR sites, usually created by reference genome annotation methods, such as the Tandem Repeats Finder (Benson, 1999). However, both methods are limited by only considering alleles that are shorter than the read length. For the Illumina HiSeq system, the typical read length is 101 nt, which covers approximately 33 repeats for a tri-nucleotide STR. Many disease-associated alleles are longer than 101 nt. For example, in Huntington's disease, the normal number of CAG repeats in *HTT* is <26, which can be covered by a read; however, the disease-associated expanded allele can reach 40 repeats, or 120 nt, thus cannot be

spanned by an Illumina read and cannot be detected by lobSTR or RepeatSeq. A similar situation occurs for spinocerebellar ataxia, dentatorubropallidoluysian atropy, myotonic distropy and other developmental disorders (Gatchel and Zoghbi, 2005).

Many medical resequencing studies adopt paired-end (PE) sequencing, in which both ends of randomly generated DNA fragments are sequenced. Typically the fragments are 300-500 nt in length, and the 101-nt sequences at both ends were determined. When an individual read fails to span a long STR allele, the read pair may still flank the STR region, thus providing additional information that can be used to detect and genotype STR alleles longer than the read length. Currently no method is available to fully extract the available information in PE data to characterize STR variation. In this chapter I describe the first algorithm to perform this task. My algorithm, *STRfinder*, is designed to detect and, when possible, estimate the length of both STR alleles using paired end next generation sequencing data. *STRfinder* does not require any prior knowledge of STR locations, and finds STR loci based on genomic distribution patterns of mapped reads or read pairs containing repetitive portions. Our algorithm is capable of finding novel STR regions that have not been documented in the human genome. *STRfinder* is implemented in Python and is available at https://sourceforge.net/projects/*strfinder*/.

4.2 *STRfinder* pipeline

4.2.1 Scope of STRfinder

In this paper, I constrain our discussion to simple repeats. This is referring to repeat region with only one type of unit, no other repeat regions at the upper or lower 300bp regions. Also the region must contain fewer than or equal to in total 2 mismatches, gaps or sequencing errors. The repeat allele with its flanking regions must appear in the genome only once. Regions violating the above standards are referred to as complex regions. Furthermore, in our context, the length of a repeating unit α takes values from 2 to 6, and the length of a repeat region has minimum number of repeat 8-α. For example, for tri-nucleotide repeat, I require at least 5 consecutive units to be called as a repetitive region.

4.2.2 Definition of STR allele types

Let $D=N\times\alpha$ denote the length of STR allele, where α is length of repeating unit and N the number of repeats. Let L_r denote read length, and for data generated from Illumina HiSeq machines, it is a non-random parameter. There are in 2 possible ranges that D falls in regarding L_r:

A $0\leq D\leq L_r-\delta$

B L_r-δ<D

δ is the minimum length of bases required for an aligner to map a read allowing soft clipping. For example, BWA requires at least 20 bases to properly map a read to the reference genome. In a region with repeats, δ is the minimum length of flanking regions. In practice, I allow δ to be user-defined, with default 20, following BWA convention.

4.2.3 Positional notations of reads and read pair types (RPTs)

Consider an STR allele with known start and end coordinates on the reference genome. I introduce positional notations for reads. In total, there are four types of reads in the region of an STR allele: at junction, inside the allele, outside the allele or traverse the allele. Let *j* denote junction read which contains repeats at either tail, *o* denote read located outside the STR region, *i* denote read inside STR region so that containing pure repeats and *t* denote traversal reads that contains repeats only in the middle and unique sequences at both ends.

Each read pair consists of two mates and I seek to denote the positions of both mates using the combination of the above notations. As a convention, I always write the left read on the left, and right read on the right. For the STR region of interest, let *L* denote the left boundary, and *R* denote the right boundary. Now I am ready to use the above notations to deliver the positional information of all the read pairs mapped within and around an STR region. For example, read pair *oLj* has left read located outside the left boundary of STR region and the right read on the junction of the starting position, and *jj* the junction-junction read pair that span the entire STR region. I enumerate and number of all the possible combinations for the above notations (**Figure 4.1**) and there are in total 13 read pair types

(RPTs) that can be generated by different allele types.

4.2.4 Characteristic RPTs for each STR allele type

Each allele type produces a subset of RPTs. It is intuitive that when $D \leq L_r - \delta$, it is impossible to produce inside reads. Therefore RPTs 1-5 will be completely missing, and all 6-13 RPTs may be observed. When $L_r - \delta < D$, the STR allele can produce RPTs 2-7 and 10-12. Traversal reads will be missing in this case. This is because even though D may be shorter than a read, the length of flanking regions left to anchor the read is not sufficient. RPT 1 may also be observed when D is long enough.

4.2.5 RPTs distribution for different allele types

I am interested in learning the behavior of STR alleles of different lengths in terms of what RPTs they can produce and what the relative proportions of different RPTs are. It is intuitive that when D is shorter than a read, it will produce a different subset of RPTs from when D is much longer than a read. To simplify our calculation, I fix insert size to be its mean μ, and assume that the left-most position of a read pair is uniformly distributed around the region $(st-\mu, st+D)$, where st is the start position of STR region and let L_t denote the length of the region: $L_t = D+\mu$. Let θ denote the length of unsequenced region between two reads in a pair, i.e. $\theta = \mu - 2L_r$. In a well-designed library, θ should be positive, although the sign of θ does not affect the results. I use M_i, $i=1,2,\ldots,13$, to denote the expected length of range where RPT i can be produced. The values of $M_{I\ am}$ given in **Table 4.1**. These values can be inferred using a series of tiling read pairs with left-most position approaching from $st-\mu$ to st+D.

Examples for $\theta<L_r$ and $D\in(\theta,L_r)$ or $D\in(\theta+L_r,\mu)$ are shown in **Figure 4.2**. And the probabilities of observing each type of RPT are given by:

$$P_1 = M_1/L_t$$

$$P_2 = P_5 = M_2/L_t$$

$$P_3 = P_4 = M_3/L_t$$

$$P_6 = P_{10} = M_6/L_t$$

$$P_8 = P_9 = M_8/L_t$$

$$P_7 = P_{11} = M_7/L_t$$

$$P_{12} = M_{12}/L_t$$

$$P_{13} = M_{13}/L_t$$

Given L_r and μ, I can calculate probabilities to observe each of the 13 RPTs for different allele length D. **Figure 4.3** displays the distributions of P_1-P_{13} with D progresses from 10 to 1000bp (μ=252, 302 and 352, L_r=101). There are several signatures revealed from **Figure 4.3**: 1) P_8+P_9, the probability of observing traversal reads, peaks when D is the smallest and decreases to 0 when D exceeds L_r. 2) $P_{13}+P_{11}+P_7+P_{12}$, the probability of observing flanking read pairs, is large when D is small and drops down to 0 when D exceeds μ. 3) $P_1+P_2+P_3+P_4+P_5$, the probability of observing RPTs with at least one inside read, remains 0 when $D<L_r$, and becomes steadily larger with increasing D. The above probability distribution provides the basis of our binary classification of the possible genotypes which I will discuss below.

4.2.6 Genotype classification for a diploid STR locus

At a STR locus in a diploid genome, there are two alleles with possibly different repeat lengths. Let L_1 denote the length of shorter allele and L_2 the longer allele, so that $L_1 \leq L_2$. Considering different ranges of L_1 and L_2, I can divide the genotypes at a given STR locus into **three scenarios**, depending on the lengths of the two alleles (L_1, L_2) , L_r:

AA: $L_1 \leq L_2 < L_r - \delta$

AB: $L_1 < L_r - \delta < L_2$

BB: $L_r - \delta < L_1 \leq L_2$

In **Table 4.2** I enumerate all the possible RPTs that can be produced under each genotype. It is intuitive that I use the existence a subset of these 13 RPTs to distinguish genotypes. Genotype AA can be separated out by missing inside reads, while AB and BB can be further distinguished by existence of traversal reads.

4.2.7 STR allele length estimation

Another major task for our method is to provide accurate estimation of L_1 and L_2 after identification of the genotype. Throughout our method development, three length estimation approaches are used for different cases:

A) exact estimation using traversal reads

B) parametric model based likelihood estimation using insert size distribution

C) non-parametric model based likelihood estimation using coverage distribution

For alleles shorter than L_r, method A) is usually sufficient to estimate D. When allele length exceeds L_r I need method B) and C) to provide unbiased estimations.

4.2.8 STRfinder pipeline

Our goal is to use previously aligned paired-end short read sequencing data to detect STR alleles without any prior information, which can be broken down into three specific tasks. The first task is to screen for informative reads and locate STR allele. *STRfinder* profiles all the reads and select partially or fully repeat ones to find candidate STR regions characterized by local cluster of partially repeat reads. Second, genotyping: using all the informative read pairs around the STR region, *STRfinder* identifies the genotypes of each STR loci. Third, length estimation: using a maximum likelihood approach, I am able to provide an estimate of the length of repeat region, even when it is longer than a read or insert size. The flowchart of *STRfinder* can be found in **Figure 4.4** and details for each task are discussed in **Method** section.

4.3 Application to simulated datasets

In order to evaluate the performance of *STRfinder*, I created *in silico* datasets with known STR allele lengths. To fully restore the complexity of human genome, without losing generality, I use chromosome 10 from human g1k v37 assembly as template. Although *STRfinder* does not rely on external information to locate an STR, I use tandem repeat table from UCSC genome browser, which was used by lobSTR to call STR variants (Gymrek et al., 2012). For chr10, there are 11,048 STR loci,

from which I randomly sampled 30% sites to be heterozygous and control the allele lengths, with the remaining 70% sites to be homozygous with reference allele length. For the 30% sites, I further sample 80% out of them (24% of total) to be unexpanded alleles, and assign one allele to be the reference. The number of repeats of the other allele is given by n_{ref}+U, where n_{ref} is the repeat content for the reference allele, and U is a random integer sampled from [-5,5]. 20% out of the sampled 30% sites (6% of total) are expanded alleles, and I still assign one to be the reference. The number of repeats of the other allele is given by U'×n_{ref}, where U' is a random integer from [2,5].

After the locations and lengths of STR alleles are chosen, I simulate reads from the simulated chr10 to obtain FASTQ files. Paired end simulated Illumina reads were generated using simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/). I set mean insert size to be 300bp, read length 101pb and mean coverage to be 50X. Other parameters are set as default. To examine the performance of *STRfinder* on different aligners, I applied two aligners that allow gap mapping, BWA (Li and Durbin, 2009) and Bowtie2 (Langmead and Salzberg, 2012) to map the reads onto human g1k v37 reference genome. I used the complete genome as reference instead of only using chr10, to include scenarios when expanded STR alleles are mapped to other chromosomes of the genome, which is possible in real datasets. In total, 3384 STR sites are selected to evaluate the performance of STR calling algorithm.

STRfinder was applied on both BWA and Bowtie2 aligned BAM files and generated two lists of STR sites. For BWA BAM, *STRfinder* called 22832 sites in total, with 22815 on chr10, 8770 in the tandem repeat table and 2924 in the heterozygous STR list, where for Bowtie2 BAM there were 22460 sites called, 22459 on chr10, 8623 in the table and 2825 belonged to the list. In either case, *STRfinder*

found more than 10K STR loci not covered in the tandem repeat table. From those loci, I randomly selected ten, manually checked their validity and all of them were proven correct STR regions.

BWA and Bowtie2 calling results overlapped 21646 in total, and 2806 were in the STR list. I compared the length estimations for both sets with the allele lengths in the STR list, and BWA result had a Pearson's correlation of 0.44, slightly larger than Bowtie2 (0.42). The two sets of BAM files yield very similar call rates and accuracies, indicating that *STRfinder* works fine with both methods. Since BWA result has slightly better outcome, I use the STR variants called from BWA aligned BAM file to compare with other STR callers.

4.4 Performance of *STRfinder* and comparison with other methods

I applied two other STR detect algorithms, lobSTR and RepeatSeq to the same dataset and compared the performances with *STRfinder*. In order to compare methods fairly, I used the optimum settings for each algorithm in our simulation framework. For lobSTR, I used paired-end mode on fastq files. If, for a given site, lobSTR provided more than one estimates, I use the one closest to the reference allele to reduce noise. For RepeatSeq, BWA aligned reads yield better call rate, and hence I used BWA instead of Bowtie2 aligned bam file. Both lobSTR and RepeatSeq required a list of reference STR sites, and I used the tandem repeat table for chr10 that has been used to generate the heterozygous STR list containing 3384 sites.

In total, lobSTR called 9410 sites, 9369 of which were on chr10, while RepeatSeq called 8859 in total and 8824 were in the tandem repeat table for chr10.

I first compare the overall call rates for three methods. Of all the 3384 sites, *STRfinder*, although is blind to the tandem repeat table used for simulation, called 2924 of them, which is the highest, followed by lobSTR, calling 2293 sites. RepeatSeq called 2249 sites. The Venn diagram in **Figure 4.5A** shows the overlaps out of two-way and three-way comparisons. 1651 sites for the called sites from all three methods are shared.

For the six genotype scenarios, the results are shown in **Figure 4.5B**. A cut-off of allele length ≤80-nt is applied for A allele. For genotype AA, all three methods have comparable high call rates and most of the calls are correct. For genotype AB, only *STRfinder* provided high fraction of correct calls. lobSTR and RepeatSeq also called around 60% of the sites, but none of them were correct. For genotypes BB, the call rates for lobSTR and RepeatSeq were close to zero, while *STRfinder* was able to correctly genotype 292 out of 536 sites for this genotype.

Genotype AA and AB are scenarios where *STRfinder* is capable of allele length estimations and I presented the comparisons using violin plots (**Figure 4.5C**). AA is the most abundant in heterozygous STR list, accounting for more than 70% of all sites. Both lobSTR and RepeatSeq were designed to predict lengths on short alleles that can be spanned by a read. As expected, allele length predictions from all three algorithms were very close to simulated values, and the median absolute prediction errors were low, with RepeatSeq being lowest (0.5 repeat units for both short and long alleles). lobSTR has the least prediction error for short allele, which is also 0.5 unit, but its prediction error (2.0 unit) for long allele is slightly worse than *STRfinder* (1.0 for both short and long alleles). Length predictions for *STRfinder* have larger variation than other methods. The sites with extreme prediction errors were manually examined and most of them were not simple repeat regions. Since *STRfinder*

does not rely on external information to genotype STR loci, its estimations on complex repeat regions suffer from low accuracy. For AB genotype, *STRfinder* have the smallest prediction errors for both long and short alleles. For these STR loci, lobSTR and RepeatSeq failed to identify the long allele, and usually assigned it with the length of reference allele, and therefore, both of them yield high prediction errors. It is also important to note that the call rate of *STRfinder* is the highest across AA and AB genotypes, where lobSTR and RepeatSeq suffered from decreased calling rate as the length of the long allele increased. Of the expansion genotype AB, I am particularly interested in tri-nucleotide repeats. There are 40 sites simulated in total, and 36 of them were called by *STRfinder*, with 33 correctly called, indicating that for tri-nucleotide repeat expansion alleles, *STRfinder* has around 80% calling accuracy, while lobSTR or RepeatSeq, not designed to detect this type of allele, could not call any of these sites correctly.

To conclude, based on an unbiased simulated dataset and comparisons among three STR callers, *STRfinder* consistently has the highest calling rate, and decent accuracies across all six genotype scenarios, where the other two callers, which were designed for only short alleles, suffered from regressed performances when allele length becomes longer.

4.5 Application to a real exome

The data I simulated using simNGS followed a protocol of whole genome sequencing preparation library. While a wide range of research is practiced using whole exome sequencing (WES) for cost efficiency considerations, I am interested to understand the performance of *STRfinder* for WES data.

The data I used came from a patient diagnosed with Machado-Joseph disease (MJD). Blood DNA sample was collected and sent for whole exome sequencing to 40X using Illumina HiSeq 2000 platform, and NimbleGen V3 capture kit. The raw paired-end reads were aligned to human reference genome g1k v37 using BWA aligner. I applied *STRfinder* onto this dataset and discovered 9791 STR sites. MJD is a neurological disorder known to be related to polygutamine (PolyQ), or CAG repeat expansion in certain genes (Paulson, 2012). I looked for pathological expansions of tri-nucleotide repeats among the 9791 sites and found seven sites with genotype AB/AC. I manually checked all the sites and found all four AB genotyped sites and the first AC genotype site located in complex interrupted repeat regions. I therefore excluded these five sites from our analysis. The sixth signal resides in chr13: 70713514-70713560, and targeted gene ATXN8OS, with short allele being 20 repeats, and long allele ~70 repeats. The last signal is in chr14: 92537353-92537396, affecting ATXN3 gene, with expanded allele length ≥70 repeat units and normal allele 20 repeats. For this specific patient, clinician has ordered Sanger sequencing on ATXN3 and validated that it has 84 repeats for the expanded allele. ATXN3 is a known causal gene for MJD, with normal allele range 13-36 repeats and pathological allele 61-84 repeats (Gatchel and Zoghbi, 2005), so the discovery of CAG expansion in this gene concluded the study. However, it will also be interesting to look at gene ATXN8OS in the future, since it has been associated with a form of spinocerebellar ataxia (Koob et al., 1999). For ATXN3 gene, lobSTR reported 13 repeats for both alleles, while RepeatSeq failed to identify the locus.

4.6 Methods

Before detailed descriptions of the method, I first specify our tasks. The goal is to use paired-end next generation sequencing data, which has been previously aligned to a reference genome, to sequentially report (1) existence of STR region; (2) the genotype and (3) lengths for L_1 and L_2.

4.6.1 Existence of STR region

4.6.1.1 Informative read searching

Our algorithm implements an initial search for all the reads that contain more than 8 repeat units at either end, to include junction reads and inside reads. To distinguish the junction read from inside reads, I perform a finer scan by searching for reads with auto-similarity larger than or equal to 0.9, i.e. if a read sequence shares more than 90% of same bases to itself lagged by α bases (iterating α from 2 to 6), it is considered to be an inside read and α is the length of repeating unit, while reads with auto-similarity smaller than 0.9 are considered to be junction reads. Each read is then paired with its mate. Each read pair filtered in contains at least one mate that is partially or fully repetitive. The above approach guarantees to find all the read pairs from RPTs 1-7,10-12 that have been assigned by aligner as unmapped, low mapping quality, soft-clipped or misplaced due to tandem repeats. For RPT 1, both mates are fully repetitive, and it is usually impossible for aligner to uniquely place the read pair in the genome, I exclude it from downstream analysis unless otherwise mentioned. RPTs 8,9 contain repeat in the middle of the sequence, and 13 does not contain repeat but spans a repeat region.

These read pairs are not included by this step and will be retrieved later. During the initial search, I also obtain the mean (μ) and standard variation (σ) of insert size distribution, using properly mapped read pairs (mapping quality≥20). To note, the thresholds used to filter in informative read pairs are adjustable by users in *STRfinder*.

4.6.1.2 Read set discovery

For each read pair kept from the above screen, I use the left-most mapping position of the mate with higher mapping quality and sort these positions along the genome for a linear scan. If in a locus I find more than 5 reads with maximum distance between adjacent reads ≤μ, I assign all these reads in this locus to be a pre read set (pRS). The discovery of a pRS is an evidence for the existence of STR region. The minimum (st) and maximum (ed) mapping locations of reads in this pRS are found and all the reads mapped within (st,ed) are retrieved to be assigned as a read set (RS). To note, RS may contain mapped read pair types 2-7,10-12, while it may also contain traversal reads 8 and 9, and traversal read pair 13, since I extract all the reads in this STR locus, and 8,9,13 read pairs are usually properly mapped.

4.6.1.3 Repeat unit identification

In the RS defined above there are two types of reads: those with decent amount of repeats (auto-similarity>=0.5) and those without (auto-similarity<0.5). I select reads with repeats to identify the repeat unit. I first estimate the length of the unit by checking the auto-similarities of these reads, where the correct length α results in the highest auto-similarity. I then enumerate all the possible α-length units, taking into account base-rotation symmetry and reverse complementary. For example,

AGC, GCA, CAG, GCT, CGT and TCG are all considered to be the same, and I take the first one by alphabetical order, AGC. For each unique repeat unit candidate, I search for its occurrence in RS, and if any read contains more than 8-α repeats for the specific unit, I include it into our repeat unit list (RUL). A clean repeat region should have only one unit, but a more complicated region, for example, containing two different yet closely located STR regions, may have more than one units. For those regions, I process one unit a time to find all the simple repeat regions within (st, ed).

4.6.1.4 STR coordinates estimation

I proceed to estimate the precise location of the STR region in the reference genome. I now focus our discussion on one unit. For region contains more than one repeat unit, same method applies for each unit iteratively. To find the precise start position (st_0) of the STR, I first find all the reads that contains unique part on the left side and repeat on the right side, namely left junction reads (LJR). To find LJR, I iterate all the reads in RS, and for each read, I identify the repeat region, allowing for at most 2 errors. Two types of errors are tolerated in *STRfinder*: 1) mismatch and 2) insertion or deletion of one base. And these errors may either be due to technical artifacts or slippage or mismatch during STR formation (Levinson and Gutman, 1987). That is, if the repeat region contains no more than two the above types of errors, it is still considered as a continuous repeat region. This procedure avoids calling shortened alleles interrupted by non-perfect matches. Similar approach applies to find right junction reads (RJR). LJR are used to estimate st_0. For each read in LJR, I estimate where tandem repeat begins by finding the left-most position (z) of the consecutive repeat region within the read ($0<z<L_r$). If the mapping position of the read is x, then $st_0=z+x$. It is possible that each LJR gives slightly different start position estimation, due to sequencing and mapping errors. I take the median of all the

start estimates to be the final st_0. Likewise, ed_0 is estimated from taking the median of all the end estimates from RJR.

4.6.2 Genotype identification

After a read set (RS) is defined and an STR region is found, I move on to identify which of the above three genotypes L_1 and L_2 belong to. It is intuitive to see in Table 2 that the missing inside reads RPT 2-5 is a signature for genotype AA, while the existence of RPT 8 or 9 further separates AB from BB. Therefore I am able to identify genotypes based on the existence of the above characteristic RPTs. Before length estimation for STR alleles, I want to describe our question. The goal is to use read length, insert size and coverage information around a diploid STR locus to provide unbiased estimation of both parental alleles, under each genotype scenario. The genotype classification of AA, AB and BB here is primitive since allele type B covers a large range of values. In allele type B, when the STR region is short enough, I expect read pairs to flank the allele (denoted as B_{short}), while its length exceeds the maximum insert size locally, no flanking read pairs shall be observed (denoted as B_{long}). However, there does not exist a definitive threshold between allele types B_{short} and B_{long}, since insert size of read pair is a stochastic variable instead of a fixed parameter. Therefore, it is not possible to define a global cut-off to separate the two possible scenarios of allele type B. Instead, I choose to analyze genotype AB or BB in a contingent fashion. For each AA genotype locus, I distinguish the two possible length ranges by looking for flanking read pairs of the longer allele. If these read pairs are found, there is definitive evidence that the longer allele is constraint by insert sizes these read pairs (B_{short}). If not, it is likely that the allele is too long to be covered by any read pair

(B_{long}). And for BB genotype, I do always not seek to provide length estimations. As mentioned in the main text, I have three methods for length estimation, with methods B) and C) for alleles longer than a read. Of these two, method B) is only available when the allele length is flanked by a sufficient number of read pairs, while method C) is applicable to either length ranges. Theoretically, the prediction error for method B) is contributed by the variance in insert size, while for method C), this error is contributed by both insert size variance and Poisson sampling error. Therefore, I use method B) for the B_{short} alleles and method C) for the B_{long} alleles.

4.6.3 Estimation of L_1 and L_2

4.6.3.1 Genotype AA: $L_1 \leq L_2 < L_r - \delta$

In this case, I expect to find traversal reads from RPTs 8,9 as well as 6,7,10-13 and no inside reads from RPTs 2-5. The method I search for *t* reads is similar to that I look for *j* reads. I iterate through all reads in RS, find reads with more than 8-α consecutive repeats in the middle, with the starting location of the repeat larger than α and ending position smaller than L_r-α, and assign them to be *t*. For each *t* read, I record the number of repeats it contains, and include it into a list (S). In this list I expect to see two numbers with high frequencies, and due to noise and errors, there may be other numbers. Therefore, I find the top two numbers with highest (f_1) and second highest frequency (f_2) in S. If $f_2 < 0.2 \times f_1$, I assume the second highest abundant number is noise and this is a homozygous region with both alleles the same length, otherwise I report heterozygous region and both numbers of repeats.

4.6.3.2 Genotype AB: $L_1 \leq L_r - \delta < L_2$

In this scenario, I expect to observe RPTs 2-13, including *t* and *i* reads. I apply method 3.a on *t* reads to estimate L_1.I then look for RPTs produced by L_2 with definitive evidence. Since L_2 is longer than L_r, it is possible to find RPTs with repetitive region longer than L_1. For example, RPTs containing junction reads (6,7,10,11,12) may fall into this category if enough number of repeats are found in one or both mates. RPTs 2-5 contain inside reads, and must be generated by L_2. I denote read pairs that can be uniquely assigned to L_2 as L_2-RP. **Figure 5.6** shows different types of L_2-RPs in different allele length ranges. For B_{short} allele type, I expect to find RPTs 7, 11 and 12 within L_2-RPs, which are very informative to L_2 estimation since they are flanking the STR region. If such read pairs exist, I use method B to estimate L_2 with following likelihood function:

$$Likelihood = \Pr\left(Read\ pair\ length\ X = L_u^i + L_2 | Read\ pair\ being\ observed\right)$$

$$= \sum_{i=1}^{K} \frac{Normal(L_u^i + L_2, \mu, \sigma)}{\int_{L_2+\delta}^{\infty} Normal(x + \delta, \mu, \sigma) dx}$$

where L^i_u is the length of non-repeat region spanned by read pair i, i=1,2…K, K is the number of L_2--RP read pairs belonging to RPTs 7, 11 or 12, and δ is the minimum length requirement for a read pair to be mapped. For BWA aligner, I choose δ=20. To calculate L^i_u, I need to know the length of the unique part of sequence coming from both mates. For RPT 12, since it travels from one end of STR region to the other, it is known that the space between mates is filled with repeats, and $L^i_u = 2 \times L_r - L_p$, where L_p is the total length of consecutive repeat sequence in the two reads. For RPT 7 and 11, $L^i_u = 2 \times L_r - L_p + G$, where G is the unique sequence between the two mates. For RPT 7, G=Start position

of right read-ed_0 and for 11, G=st-Start position of the left read-L_r. The denominator in the likelihood formula is the probability that the read pair is observed. Conditioning on the read pair observed and belonging to L_2, the insert size follows a truncated normal distribution where X has to be greater than $L_2+\delta$, and given L_2, the probability is integral of normal density from $L_2+\delta$ to infinity. I numerically solve L_2 to maximize likelihood function and report genotype AB_{short} as well as both length estimations.

When RPT 7,11 or 12 within L_2-RP are missing, it is likely that L_2 belong to allele type B_{long}. I use method C, i.e. coverage based likelihood model, to estimate L_2. In practice, I find all the reads that are known to be produced by L_2, knowing that $L_1<L_r$. It is straightforward that RPTs 2-5 are all L_2 generated as well as RPT 6 or 10 containing more repeat units than the shorter allele. Similarly, RPT 8 and 9 must be produced by L_1, since they traverse the shorter allele. I let N_{long} denote the number of read pairs produced by L_2, which is summation of the counts of L_2-RPs, and N_{short} denote the number of read pairs produced by L_1, which is the count of RPT 8 and 9. Let P_{long} and P_{short} denote the probabilities that I can observe an L_2 read pair and L_1 read pair respectively. **Figure 4.6** shows details of P_{long} and P_{short} calculations. Both probabilities need to be mathematically inferred. To simplify the calculation, I let insert size to be constant μ. Follow the above discussion, the chance that a read pair can be observed as L_2-generated is:

$$p_{long}$$

$$= \Pr(read\ pair\ truly\ comes\ from\ L_2\ and\ contains\ enough\ repeats\ to\ be\ assigned\ to\ L_2)$$

$$= \frac{L_2 + \square}{2\times\mu + L_1 + L_2} \times \frac{L_2 + \mu - 2\times L_1}{L_2 + \mu} = \frac{L_2 + \mu - 2\times L_1}{2\times\mu + L_1 + L_2}$$

where the first part is the probability that a read is truly generated by L_2, and the second part is the

probability that the repeat contain in the read pair is longer than L_1 so that it can be assigned to L_2 definitively. Likewise, the chance that a read pair can be observed as L_1-generated is:

$$p_{short}$$

$$= \Pr(read\ pair\ truly\ comes\ from\ L_1\ and\ contains\ enough\ repeats\ to\ be\ assigned\ to\ L_1)$$

$$= \Pr\ (observe\ traversal\ read) = \frac{L_1+\mu}{2\times\mu+L_1+L_2}\times\frac{2\times(L_1-L_r)}{L_1+\mu} = \frac{2\times(L_1-L_r)}{2\times\mu+L_1+L_2}$$

Let N_{tot} denote the number of read pairs in RS and it follows $N_{tot}>N_{long}+N_{short}$ and $P_{long}+P_{short}<1$, since there are a fraction of read pairs cannot be assigned to either L_1 or L_2. I build a Multinormial likelihood model based on coverage:

$$Likelihood = Multinomial(p_{short}, p_{long}, 1-p_{short}-p_{long}, N_{short}, N_{long}, N_{tot}-N_{short}-N_{long})$$

It can be shown that the maximum likelihood estimation for L_2 is unbiased.

I apply this model, obtain an MLE estimation $\widetilde{L_2}$ and report genotype AB_{long} and both length estimations for this situation.

4.6.3.3 Genotype BB: $L_r<L_1\leq L_2$

In this scenario, I do not expect to observe any RPT 8 or 9.There are two possibilities: i) flanking read pairs RPT 7,11 or 12 can be observed ($B_{short}B_{short}$ or $B_{short}B_{long}$) and ii) no flanking read pairs observed ($B_{long}B_{long}$). It is not possible to assign these read pairs to be L_1-reads or L_2-reads as I did for 3.b. But for i), it is still possible to provide some information of allele lengths. I first distinguish two cases: flanking read pairs were generated from (1) one allele ($B_{short}B_{short}$ homozygous or $B_{short}B_{long}$) or (2) two alleles ($B_{short}B_{short}$ heterozygous). For (1), I expect to observe the variance of insert size of flanking read pairs to be very close to σ. For (2), the flanking read pairs actually come from two

distributions with different means, and I expect the observed variance σ' to be greater than σ. In practice, if $\sigma' \geq 1.2\sigma$, I assert case (2) and report genotype $B_{short}B_{short}$ heterozygous. If σ' is comparable to σ, two scenarios may be true: 1) $B_{short}B_{long}$: the longer allele is too long to be covered by flanking read pairs, and the observed RPT 7, 11, 12 were all generated from the shorter allele and 2) $B_{short}B_{short}$ homozygous: the longer and shorter allele are of same length, or $L_1=L_2$. I use a ratio (k) of counts for RPT 7, 11, 12 over counts for RPT 6 and 10. The expected value of k is $(\mu - L_1)/2L_r$ for scenario 2), and half of that value for scenario 1). Since L_1 has not been estimated, I use $L_{ref}=ed_0-st_0$ instead in the above expression. In practice, if the observed value ($\tilde{k}$) is greater than the 0.8×Exp[k], I assign genotype $B_{short}B_{short}$ homozygous, and I test if L_1 is the reference allele length by using one-sample student t-test of mapping distance against μ. If P-value ≤0.05, then L_1 is estimated to be $L_{ref}+\Delta$, where Δ=μ-average insert size of flanking read pairs in the locus. If P-value >0.05, $L_1=L_2=L_{ref}$. I report $L_r<L_1=L_2<\mu$ for either P-value<=0.05 or P-value>0.05, together with the length estimations. If $\tilde{k} \leq \frac{5}{8}k$, I assign genotype $B_{short}B_{long}$ and the insert sizes of flanking reads come from one normal distribution. I directly estimate L_1 using method B, described in 3.b and report the genotype and the estimation for L_1 only.

If no flanking read pairs were observed, I do not seek to make quantitative estimations for either L_1 or L_2 and will just report the length ranges and genotype $B_{long}B_{long}$.

4.7 Summary

I present *STRfinder* as a new approach to detect, and when possible, to genotype short tandem allele loci in genomes using short read DNA sequencing data. Compared with previously developed methods, *STRfinder* has several advantages: 1) it does not rely on a known list of STR loci to call or genotype STRs. Therefore, *STRfinder* is fully capable of finding true de novo STR sites; 2) it integrates all 13 read pair types around a repeat region to robustly call and genotype STR alleles; 3) for a wide range of STR allele length, *STRfinder* provides accurate estimations with high call rates. These features make *STRfinder* a unique tool to detect pathological tandem repeat expansions and find the genetic basis of a wide spectrum of neurodevelopmental diseases.

To date, in our simulation, *STRfinder* called fewer sites in the tandem repeat table than lobSTR or RepeatSeq. This is because in our application, I used a more stringent criterion to call repeats, that the allele must contain more than 8 repeats. If I relax this threshold to 6 repeats, *STRfinder* called 10120 sites in the tandem repeat table, which is more than either lobSTR or RepeatSeq. In practice, this threshold is user-defined.

STRfinder estimates alleles longer than a read by using the mapping distance distribution of the library. I assume it follows a single mode Gaussian distribution. The empirical distribution can be used instead to increase accuracy. Larger library size is preferred to detect longer alleles. For example, with median insert size 300-nt, it is possible to estimate repeat allele length up to 260-nt, considering 20-nt is required to anchor the junction reads, which is enough to cover the expanded alleles of Huntington's disease, or various types of Ataxia disorder.

Application of *STRfinder* on *in silico* generated dataset with known number of repeats for a list of selected STR loci showed that most sites were accurately genotyped by *STRfinder*. Comparing with

lobSTR and RepeatSeq, *STRfinder* provides equally good estimations for shorter alleles with additional information for expanded alleles, both their genotypes and length estimations. Although *STRfinder* does not perform realignment of repeat reads as lobSTR does, its performance is not suffered, since a comprehensive collection of read pair types are utilized around a STR region to increase calling accuracy. BWA and Bowtie2 aligners were tested in our simulations and both provided similar results. Although I did not test other aligners that support mapping of small insertions and deletions, by design they are all valid options to apply *STRfinder*.

It took *STRfinder* 2.8 CPU hour to process the simulated dataset (AMD Opteron 2.4GHz), and approximately 5GB memory consumption. The actual memory usage depends on the size of the BAM file and parameter settings. More sensitive settings result in high memory consumption. The simulated dataset contained 90 million reads for chr10 only, and *STRfinder* called ~23K STR sites. On average, it takes *STRfinder* 0.43s to fully analyze one site. Multi-thread processing is available for *STRfinder* to call variants on multiple samples in parallel.

I applied *STRfinder* to a patient with Ataxia symptom on the whole exome sequencing data, and independently discovered the CAG repeat expansion in ATXN3 gene, which had been clinically tested and proven expanded. This result indicates that our algorithm is ready for the detection of causal STR alleles in susceptible population with NGS data available. I based our methodology development on simple repeats, yet since I do not rely on known STR sites as reference, sometimes *STRfinder* will find complex repeat regions and report incorrect results, which is the major contribution of prediction errors in our simulation. Future work needs to be done to solve complex repeat regions, and to achieve this goal, a DNA fragment library with a wide spectrum of length

distribution is helpful to detect long regions with subtle structures.

Table 4.1. Distributions of lengths of range where RPTs can be produced.

A: $\theta<L_r$

	$(0,\theta)$	(θ,L_r)	$(L_r,\theta+L_r)$	$(\theta+L_r,\mu)$	(μ,∞)
M_1	0	0	0	0	$L-\mu$
M_2,M_5	0	0	$L-L_r$	θ	θ
M_3,M_4	0	0	0	$L-L_r-\theta$	L_r
M_6,M_{10}	L	L	L_r	L_r	L_r
M_7,M_{11}	L	θ	$L_r-L+\theta$	0	0
M_8,M_9	L_r-L	L_r-L	0	0	0
M_{12}	0	$L-\theta$	$L-\theta$	$L_r-(L-L_r-\theta)$	0
M_{13}	$\theta-L$	0	0	0	0

B: $L_r<\theta<\mu$

	$(0,L_r)$	(L_r,θ)	$(\theta,\theta+L_r)$	$(\theta+L_r,\mu)$	(μ,∞)
M_1	0	0	0	0	$L-\mu$
M_2,M_5	0	$L-L_r$	$L-L_r$	θ	θ
M_3,M_4	0	0	0	$L-L_r-\theta$	L_r
M_6,M_{10}	L	L_r	L_r	L_r	L_r
M_7,M_{11}	L	L_r	$L_r-L+\theta$	0	0
M_8,M_9	L_r-L	0	0	0	0
M_{12}	0	0	$L-\theta$	$L_r-(L-L_r-\theta)$	0
M_{13}	$\theta-L$	$\theta-L$	0	0	0

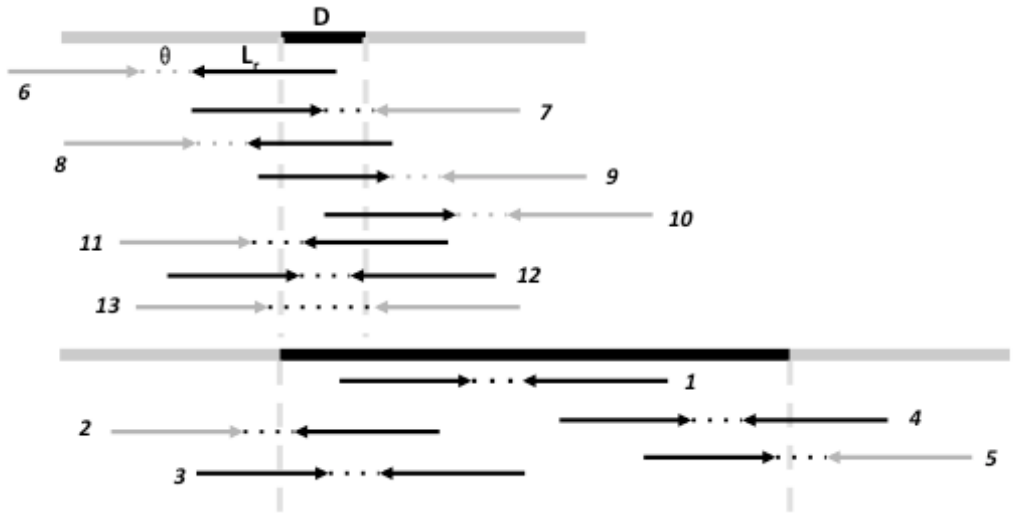
Upper table is for case when θ is shorter than a read and the lower table is when θ is greater than a read. It is impossible for θ to be larger than μ, since read length is non-negative.

Table 4.2: RPT distributions across six genotypes and binary classification of 6 genotypes.

		Notation	AA	AB	BB
1		*ii**	••	••/•√	••/•√/√√
2		*oi*	••	•√	√√
3		*ji*	••	•√	√√
4		*ij*	••	•√	√√
5		*io*	••	•√	√√
6		*oLj*	√√	√√	√√
7		*Ljo*	√√	•√/√√	••/•√/√√
8		*ot*	√√	√•	••
9		*to*	√√	√•	••
10		*Rjo*	√√	√√	√√
11		*oRj*	√√	•√/√√	••/•√/√√
12		*jj*	√√	•√/√√	••/•√/√√
13		*oo*	√√	•√/√√	••/•√/√√

•: missing, √: presented, *: RPT1 is unmappable. Existence of RPT 8 and 9 (shaded gray) are used to separate AA,AB and AC from the other three. Within the first 3 genotypes, RPT 2-5 (shaded orange) are used to further split AA out. Within the last 3 genotypes, RPT 7,11,12,13 (shaded magenta) are used to further split CC out. The second column display color legend for each RPT matching **Figure 4.3**.

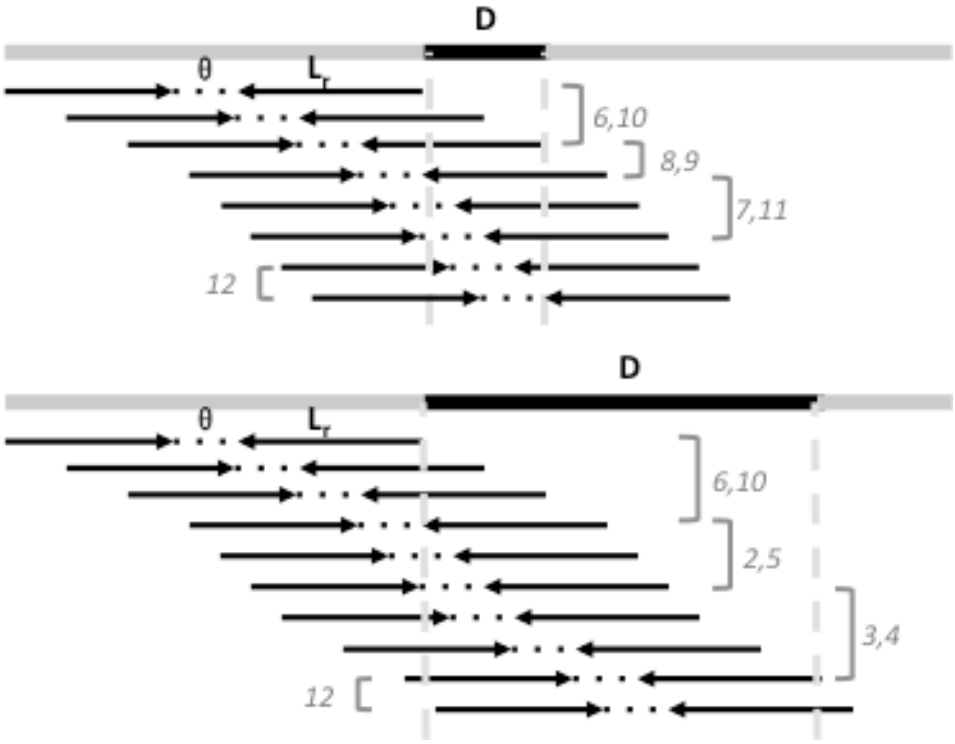
Figure 4.1: Thirteen read pair types relevant for STR detection.



RPT shortcut number	1	2	3	4	5	6	7	8	9	10	11	12	13
Positional Notation	*ii'*	*oi*	*Lji*	*iRj*	*io*	*oLj*	*Ljo*	*ot*	*to*	*Rjo*	*oRj*	*jj*	*oo*

Solid thick lines in the middle indicates diploid genomic region with STR (black colored). Black color indicates reference genome, reads or regions containing no repeats, while gray color for sequences that are at least partial repeats. The upper allele is the longer allele. Paired arrows indicate read pairs. Details for the longer and shorter alleles and the RPTs they may produce are discussed in the main text.

Figure 4.2: Cartoon showing RPT fractions as in Table 1.



Different RPTs generated when A: $\theta<D<L_r$ and B: $L_r<D<\mu$ and their relative fractions as indicated by the tiling reads.

Figure 4.3: Distribution of observing the 13 RPTs under different insert size and read length configurations, with allele length changing from 10 to 1000bp. RPTs with same probability are put next to each other. θ=μ-2L$_r$, is the between read distance in a pair.

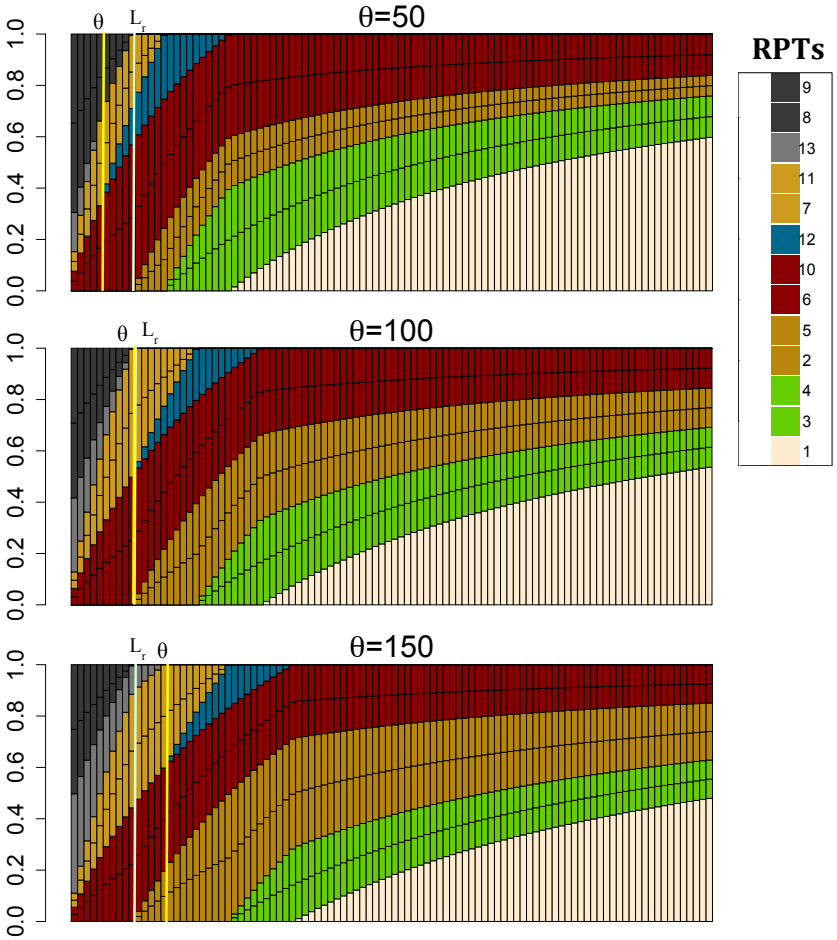
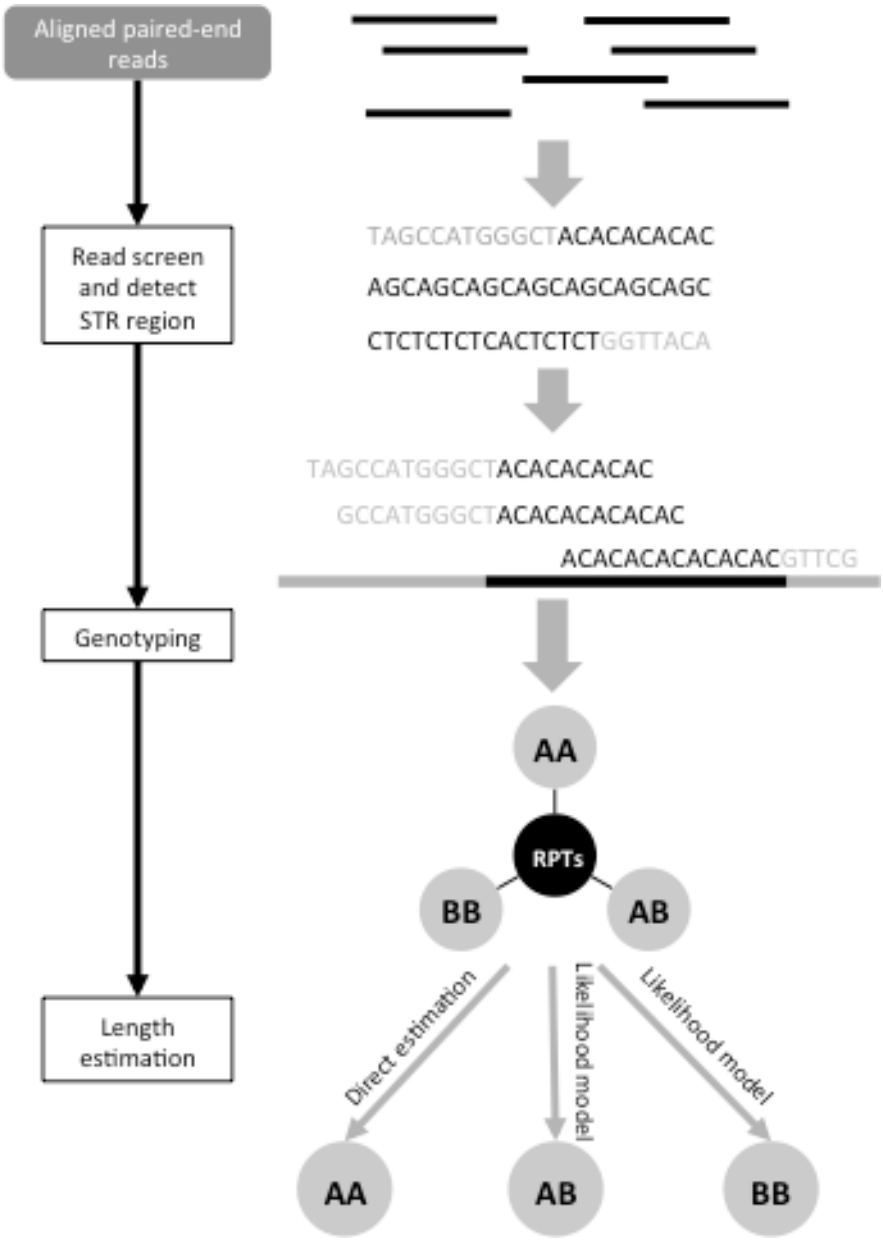
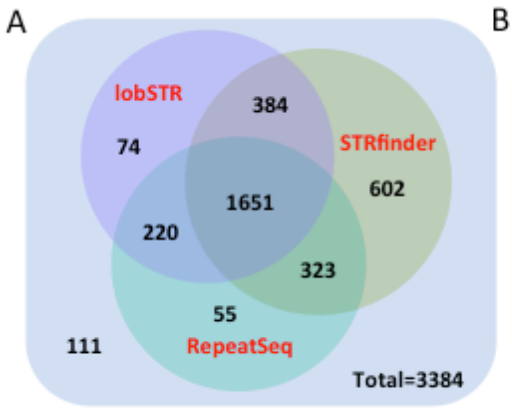


Figure 4.4: *STRfinder* pipeline



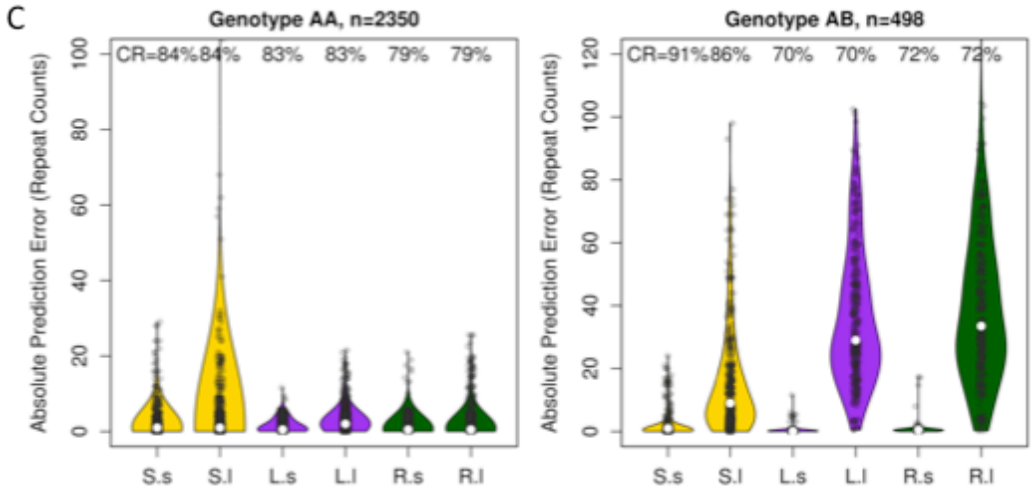
The *STRfinder* algorithm contains three steps. Assuming paired end reads have been aligned to the reference genome using indel tolerated aligners, such as BWA or bowtie2, *STRfinder* screens for repetitive reads, find read sets that contain closely located repeat reads, genotype the STR locus using informative read pair types (RPTs) and estimates the lengths of alleles when possible.

Figure 4.5: Performances comparison of *STRfinder*, lobSTR and RepeatSeq on simulated dataset.



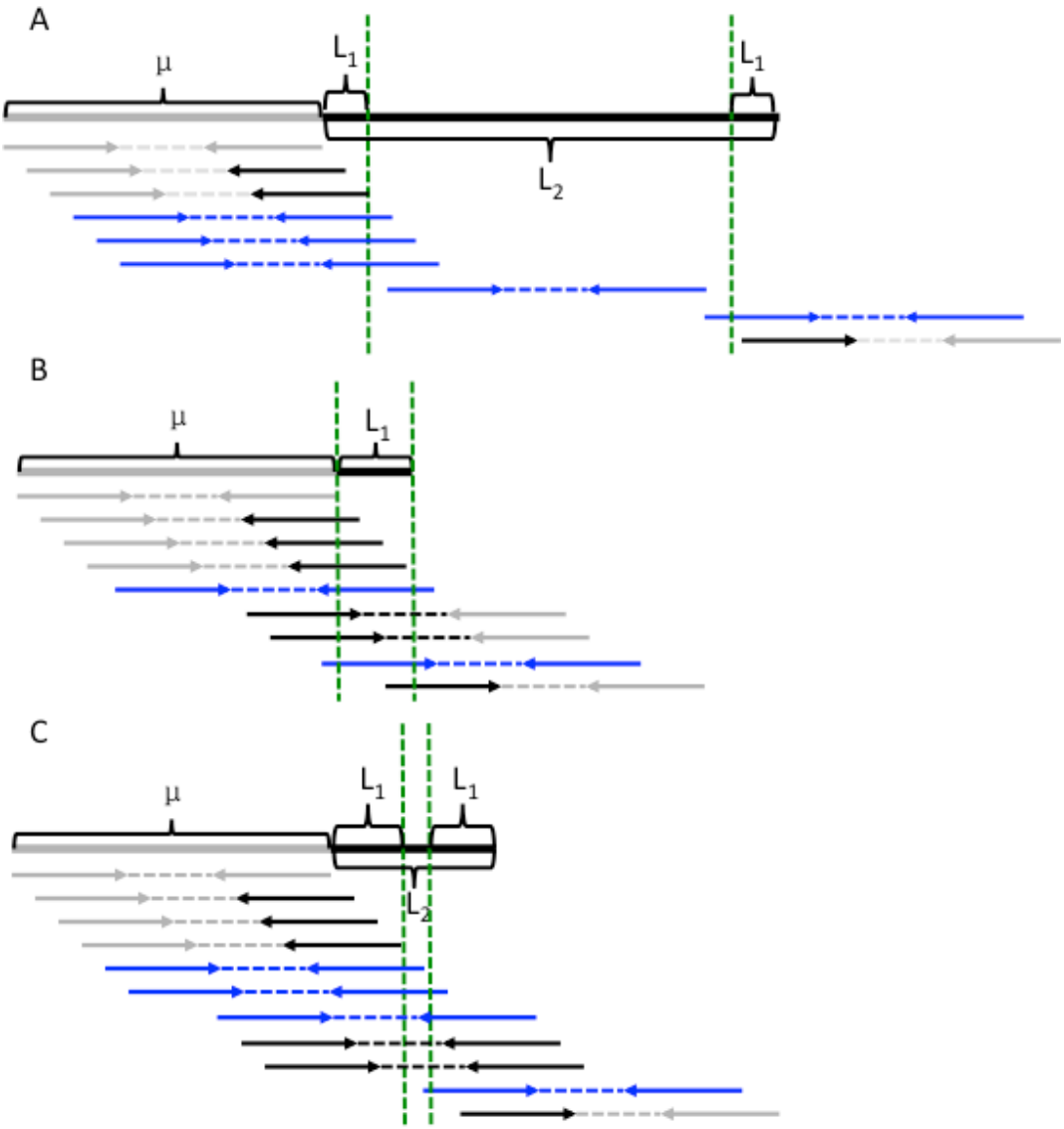
B

	n	Software	#call	#correct
AA	2350	*STRfinder*	1999	1980
		lobSTR	1941	1941
		RepeatSeq	1860	1857
AB	498	*STRfinder*	472	408
		lobSTR	347	1
		RepeatSeq	360	0
BB	536	*STRfinder*	453	292
		lobSTR	4	0
		RepeatSeq	29	2



A. Venn diagram showing the two-way and three-way overlaps between STR sites called by different callers. B. Table showing call rates and number of correctly genotyped sites by each method, for genotypes: AA, AB or AC, BB or BC and CC. C. Violin plots of absolute prediction error in unit of repeats for each method and stratified by genotype AA, AB and AC. The X-axis of the plot displays the methods and allele. Capital letters are abbreviations for STR callers, "S" for *STRfinder*, "L" for lobSTR and "R" for RepeatSeq, where the following small number designates allele: "s" for short allele and "l" for long allele.

Figure 4.6: Cartoon demonstration of P_{long} and P_{short} calculation in different scenarios



A. $\mu<L_2$: To estimate P_{long}, consider tiling read pairs with left-most position moving from st_o-μ to st_0+L_2, L_2-reads can be observed in the region of st_0-μ+L_1 to $st_0+L_2-L_1$. Assuming read pairs are uniformly distributed in the region, the probability to observe L_2-reads is $(L_2+\mu-2L_1)/(L_2+\mu)$. B. $L_1<L_r$: To estimate $P_{short,}$ read pairs with left-most position moving from st_0-μ to st_0+L_1 can be observed as RPT 8 or 9 in the region from st_0-μ+L_1 to st_0-μ+L_r and $st_0+L_1-L_r$ to st_0. The probability to observe L_1-reads is $2(L_r-L_1)/(L_1+\mu)$. C. $L_2<\mu$ and L_1 is long enough to prevent the observation of L_2 flanking read pairs. L_2 reads can be observed in the range of st_0-μ+L_1 to st_0-μ+$L_r+L_2-L_1$ and $st_0+L_1-L_r$ to $st_0+L_2-L_1$. Therefore, the probability to observe L_2-reads is $2(L_r+L_2-2L_1)/(L_2+\mu)$. L_2-reads are marked with blue color.

4.8 Bibliography

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573-580.

Bolton KA, Ross JP, Grice DM, Bowden NA, Holliday EG, Avery-Kiejda KA, Scott RJ. 2013. STaRRRT: a table of short tandem repeats in regulatory regions of the human genome. *BMC genomics* **14**: 795.

Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A et al. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**(5254): 1423-1427.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature reviews Genetics* **6**(10): 743-755.

Ge J, Eisenberg A, Budowle B. 2012. Developing criteria and data to determine best options for expanding the core CODIS loci. *Investigative genetics* **3**: 1.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics* **44**: 445-477.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome research* **22**(6): 1154-1162.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic acids research* **41**(1): e32.

Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, Day JW, Ranum LP. 1999. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature genetics* **21**(4): 379-384.

Kozlowski P, de Mezer M, Krzyzosiak WJ. 2010. Trinucleotide repeats in human genome and exome. *Nucleic acids research* **38**(12): 4027-4039.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution* **4**(3): 203-221.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

MacLean HE, Brown RW, Beilin J, Warne GL, Zajac JD. 2004. Increased frequency of long androgen receptor CAG repeats in male breast cancers. *Breast cancer research and treatment* **88**(3): 239-246.

Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW, Vogelstein B et al. 1995. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268**(5215): 1336-1338.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* **447**(7147): 932-940.

Nelson KA, Witte JS. 2002. Androgen receptor CAG repeats and prostate cancer. *American journal of*

epidemiology **155**(10): 883-890.

Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LP, Zoghbi HY. 1993. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature genetics* **4**(3): 221-226.

Parsons R, Myeroff LL, Liu B, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B. 1995. Microsatellite instability and mutations of the transforming growth factor beta type II receptor gene in colorectal cancer. *Cancer research* **55**(23): 5548-5550.

Paulson H. 2012. Machado-Joseph disease/spinocerebellar ataxia type 3. *Handbook of clinical neurology* **103**: 437-449.

Pearson CE, Nichol Edamura K, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. *Nature reviews Genetics* **6**(10): 729-742.

Popat S, Hubner R, Houlston RS. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **23**(3): 609-618.

Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A et al. 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nature genetics* **14**(3): 269-276.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one* **8**(2): e54710.

Smit AFA, Hubley, R. and Green, P. 1996-2004. RepeatMasker Open-3.0.

Walker FO. 2007. Huntington's Disease. *Seminars in neurology* **27**(2): 143-150.

Wooster R, Cleton-Jansen AM, Collins N, Mangion J, Cornelis RS, Cooper CS, Gusterson BA, Ponder BA, von Deimling A, Wiestler OD et al. 1994. Instability of short tandem repeats (microsatellites) in human cancers. *Nature genetics* **6**(2): 152-156.

Chapter 5. Conclusion and Future Directions

5.1 Conclusions

In this dissertation, I have developed a collection of biostatistical and bioinformatics tools to study intra-tumor heterogeneity and the evolution of cancer genomes. Overall, I have completed three projects under this theme.

In my first project, I studied the euploid cell mixing ratios in a cohort of human GBM samples, using allele-specific SNP array data. I discovered a strong correlation between AGP, the fraction of aneuploidy cells in a tumor sample, and gene expression PC1 and PC2, indicating that major components of gene expression variation of GBM samples are influenced by their levels of normal cell admixture. This is a novel finding, one that was ignored in the initial analysis and reporting of TCGA data. With this knowledge, I performed a joint analysis on copy number alteration profiles and gene expression differences of these samples, and revised the classification of GBM. The new subtypes are more strongly associated with patients' survival than the previously defined subtypes. Furthermore, by comparing with known neural cell types, I identified that the euploid cells in the Mesenchymal subtype are likely to be infiltrating microglia/macrophage.

In the second project, I extended my algorithm developed in the first study from estimating whole-genome average mixing ratios to the mixing ratios of individual CNAs. Application of this algorithm to a breast cancer cohort revealed that about half of the samples consist of more than one subclone. I further integrated DNA sequencing data into my analysis. With a model that considers all possible temporal orders and phase relationships between a somatic mutation and an sCNA it resides in, I inferred the cancer cell fraction (CCF) for each somatic mutation. The collection of tools, named Clonal Heterogeneity Analysis Tool (*CHAT*), is one the few methods in the field that analyze both sCNA and somatic mutations, and estimate cellular frequencies for both types of variants. It is more general than other methods by considering the widest range of possible evolutionary scenarios.

Throughout the development of this method, I relied on the regional two-way mixing hypothesis, first brought up in oncoSNP (Yau et al., 2010). This hypothesis is equivalent to the infinite site assumption used in population genetics, which considers recurrent mutation in the same locus of the genome as extremely unlikely. For somatic mutations in most tumor samples, this assumption is reasonable due to the large size of human genome and the relatively low rate of single nucleotide replacement. However, tumors with large fractions of genome altered are likely to have some regions affected by more than one independent events. And the exact solution is to allow three-way mixings or higher. While two-way mixing model is solvable using allele-specific copy number data, higher-order mixing models are mathematically difficult, sometimes becoming intractable when the data are limited. In my current approach, if an sCNA does not follow this assumption, its sAGP value will be

assigned with a missing value, and no downstream analysis will be conducted for this sCNA. My third project is to detect and genotype STR loci. While it was initially motivated to study pathological STR expansions in neurological and developmental disorders, it can also be applied to study STR variation in the cancer genomes. Compared to lobSTR and RepeatSeq, the advantage of my tool, *STRfinder* is its capability to detect and genotype alleles that are longer than the read length, making fuller use of the information contained in paired-end short read sequencing data. Furthermore, unlike lobSTR or RepeatSeq, *STRfinder* can detect novel STR regions, without the need of an existing collection of known STRs in the genome. Together, these features make *STRfinder* a valuable new addition among tools to study STR variation. When applied in studies of cancer genomes, it is expected to enhance our ability to examine microsatellite instability and its role in tumor progression.

5.2 Future Directions

There are several continuations for the intra-tumor heterogeneity project. Methodologically, the tool developed in my thesis, *CHAT*, estimates cellular frequencies for individual somatic events, but without borrowing information from other somatic events. Other approaches, including *PyClone* and *THetA,* jointly use all the variants to infer the subclonal structure, which has the advantage of reducing the noise of individual estimates, but has the drawback of forcing the somatic events in the tumor into different subclones even when the true population is uniform. An important direction of future development is to appropriately

incorporate other sites of the genome but without imposing an arbitrary tumor subclonal model in the inference, and this is expected to increase the accuracy of sAGP and CCF predictions.

Translational research that bridges fundamental biological discoveries and clinical application is an important area in biomedical science. There are several future directions for intra-tumor heterogeneity studies in this field. First, the field would benefit from the development of cost-effective validation of subclonal driver lesions discovered using high-throughput technologies. These events usually have low prevalence in the population, and it requires ultra-deep sequencing or single cell profiling to prove their existence. Second, functional analysis of subclonal driver events is desirable and model organisms or cell lines that recapitulate the hallmarks of *in vivo* cancers are in need.

Besides validations for variant discoveries, intra-tumor heterogeneity studies have close connections with clinical practice. In **Chapter 2**, I have studied the impact of tumor/normal mixing in tumor samples, and discovered that it has significant impact over tumor subtype classification. And in **Chapter 3**, I discovered that in breast tumors the clonal patterns for tumor related genes show difference across tumor subtypes. In both studies, heterogeneity within a tumor cell population influences the observation of inter-tumor diversity. In the future, discovery of clinically related tumor subtypes would benefit from accounting for intra-tumor heterogeneity characteristics, such as tumor/normal mixing ratios and cellular frequencies for somatic events.

It has been shown that the diversity in tumor cell population is a predictor for clinical

outcome (Maley et al., 2006) in esophageal adenocarcinoma patients. While in that study, researchers applied the labor intensive karyotyping to profile single cells and report subclonlaity, it is possible to apply *CHAT* on genomics data collected from bulk tissue and estimate clonal diversity in a high-throughput manner. In the future it will be particularly interesting to study the impact of clonal diversity on patient survival for more cancer types and subtypes.

As mentioned in **Chapter 1**, cytotoxic therapies rarely eradicate tumor cells, and most deaths caused by cancer are due to recurrence or metastases. Intra-tumor heterogeneity has played an important role in the relapse by providing multiple subclones carrying different somatic events, and the treatment is likely to fail if at least one of the subclone harbors drug resistant mutations. Research studying leukemia or lymphoma monitor subclonal dynamics by longitudinal sampling greatly helped to understand the evolution of tumor subclones. For solid tumors, it will also be interesting to compare the subclonal architectures in primary tumors with those in recurrent or metastatic tumors. The field is in need to understand subclonal evolution and replacement under various treatment options, to minimize the risk of relapse and optimize the clinical outcome.

In **Chapter 4** I developed *STRfinder*, where I detected and genotyped STR alleles using a subset of informative read pair types. In the future, *STRfinder* will benefit from an exact likelihood formulation that uses all the information available for the STR locus. The prediction accuracy is expected to improve with a full likelihood model that includes the coverage distribution of all 13 read pair types, their insert size distribution, and the split reads

mapping. Also, many cancer types are associated with microsatellite instability (MSI), such as GBM (TCGA, 2008), colorectal cancer (Boland and Goel, 2010), gastric cancer (Halling et al., 1999), melanoma (Kroiss et al., 2001), etc. It will be interesting to apply *STRfinder* to these datasets to identify novel microsatellite alleles that cannot be detected using previous methods.

STRfinder provides poor predictions for regions with complex repeats, which is a major limitation for the current method. And these regions can be highly interesting. For example, the recently developed technology using clustered regularly interspaced palindromic repeats (CRISPR) provides an elegant system to perform gene editing (Wang et al., 2014, Hsu et al., 2014). And it will be helpful to detect CRISPR loci in bacteria strains whose genome have not been fully assembled, using short read DNA sequencing data. To detect the closely located interspaced repeats using NGS data is a new bioinformatics challenge and currently no software is capable of this task.

5.3 Closing remarks

Since the first cancer genome was sequenced in 2008 (Ley et al., 2008), the paradigm of cancer research has been shifted. Cancer genomics data have been collected worldwide with an unprecedented speed, resolution and scale. Large cancer cohort studies have strongly shaped the landscape of cancer research, with a profound influence over cancer diagnosis, patient care, and drug development. Great opportunities have come up with an abundance of

cancer datasets, spanning a wide spectrum of cancer types and multiple levels of biological regulations, yet along emerged new challenges. One of the greatest challenges is to identify driver mutations required for rapid malignancy growth and expansion. Algorithms developed in my thesis provide methodological support to quantitatively evaluate the effect of somatic events in the context of clonal structure. I also presented a novel approach to discover and genotype an understudied type of genetic variation, short tandem repeats, or microsatellites, and this tool will be helpful to understand tumor genome evolution. It is foreseeable that more genomics data derived from bulk tumor tissue will be collected for a wider range of cancer patients and the methods developed in this dissertation are expected to serve the need of analyzing and interpreting cancer genomics data.

5.4 Bibliography

Boland, C. R. & Goel, A. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology,* 138**,** 2073-2087 e3.

Halling, K. C., Harper, J., Moskaluk, C. A., Thibodeau, S. N., Petroni, G. R., Yustein, A. S., Tosi, P., Minacci, C., Roviello, F., Piva, P.*, et al.* 1999. Origin of microsatellite instability in gastric cancer. *Am J Pathol,* 155**,** 205-11.

Hsu, P. D., Lander, E. S. & Zhang, F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell,* 157**,** 1262-78.

Kroiss, M. M., Vogt, T. M., Schlegel, J., Landthaler, M. & Stolz, W. 2001. Microsatellite instability in malignant melanomas. *Acta Derm Venereol,* 81**,** 242-5.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., Mclellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., Mcgrath, S., Hickenbotham, M.*, et al.* 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature,* 456**,** 66-72.

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R. A., Rabinovitch, P. S.*, et al.* 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet,* 38**,** 468-73.

TCGA 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455**,** 1061-8.

Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science,* 343**,** 80-4.

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O. & Holmes, C. C. 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol,* 11**,** R92.