Bayesian Functional Data Methods in Copy Number Alteration Studies: Applications in Urothelial Bladder Carcinoma

Miranda L. Lynch* Jessica M. Clement†

Abstract

Chromosomal level alterations in the genome are a hallmark of cancer, and methods to probe copy number alterations (CNAs) have revealed a number of important drivers of oncogenesis and tumor progression. This work is motivated by our research questions in urothelial bladder carcinoma (BLCA), investigating the interconnections between smoking status/history and immune response in BLCA progression. In this work, copy number profiles are characterized using Bayesian functional data methods employing wavelet basis functions. These basis functions are well suited for the types of profiles that appear in copy number studies using array CGH and SNP arrays. We propose methods using these profiles for functional regression to examine the relationship to smoking and to look at whether the altered genomic regions preferentially include genes indicative of immune response. We apply our methods to publicly available bladder cancer data from a group of patients with metastatic disease.

Key Words: Bayesian inference, functional data regression, bladder cancer, copy number alterations

1. Introduction and Motivating Problem

This applied work focuses on making use of functional data analysis methods applied to high dimensional copy number data to address questions in urothelial bladder carcinoma (BLCA). While copy number alterations (CNAs) have been implicated in cancer, it remains unclear whether the observed alterations represent general genomic instability that is characteristic of tumor progression, or if the alterations signal key driver events (amplification of oncogenes or deletion of tumor suppressor genes), or even have a role in the process of tumorigenesis itself [5, 18, 2, 20]. As such, appropriately characterizing copy number alteration events is an important step in both delineating the role of large-scale structural genomic alterations in cancer, and in highlighting novel tumor suppressors and oncogenes. The goal of the present work is to develop methods specific to the nature of copy number measurement data, working in a regression context in order to address questions about patient characteristics associated with certain copy number patterns. While most CNA analyses have focused on locating and cataloguing genomic loci of CNA, this work focuses on using functional data regression methods to compare how patient traits associate with different CN profile patterns. In particular, in the arena of urothelial bladder carcinoma, we are interested in comparing copy number profiles in patients with muscle-invasive bladder cancer (MIBC) versus non-MIBC patients, and to correlate whether CN losses/gains associate with given immune response genes. To this end, here we develop functional data regression methods using wavelet basis functions that address common modes of copy number measurement.

*Center for Quantitative Medicine, University of Connecticut Health Center, 263 Farmington Ave MC 6033, Farmington CT 06030

†University of Connecticut Neag Comprehensive Cancer Center and Dept. of Hematology, 263 Farmington Ave, Farmington CT 06030

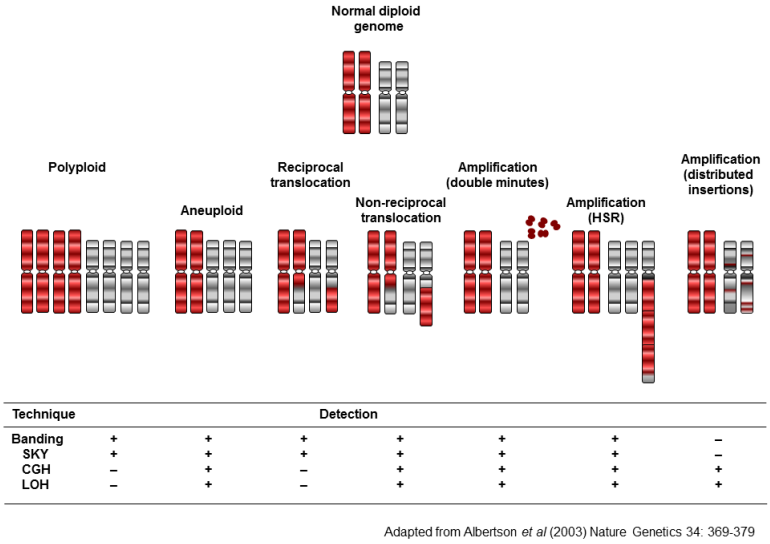

Figure 1: Schematic depiction of different mechanisms and a listing of some of the techniques used to detect chromosome-level aberrations. Adapted from [2].

1.1 Copy number data

The area of application for the functional data regression methods under consideration is copy number data. CNAs are large, structural genomic alterations, typically ranging in size from $\sim$100 kb to several mega-bases (Mbs), and can include very large events such as deletion/amplification of whole chromosome arms, and duplication or deletion of entire chromosomes. In the context of cancer biology, CNAs typically refer to somatic genomic alterations, as opposed to germline variants. CNAs are detected as regional deviations from the normal diploid condition of genomic material. A schematic of types of genomic alteration, and listing of which CN measurement modalities are capable of detecting the different types of variants, appears in Figure 1, adapted from Albertson, 2003 [2].

Several technical platforms have emerged for detecting large scale structural genomic alterations in cancer tumor samples (reviewed in [3]). These include methods for whole-chromosome visualization (such as FISH, flourescence *in situ* hybridization), hybridization methods that query locations along the entire length of the genome, including SNP arrays and array comparative genomic hybridization (CGH) methods, and sequencing-based characterization of genomic content (such as RNA-seq) [9]. In this work, we focus on the data structure arising from array CGH methods.

Array CGH is a popular method for examining CNAs, available on many commercial platforms. In CGH methods, DNA isolated from the tissue of interest (eg tumor tissue) is differentially labeled relative to a reference sample (eg normal tissue), and then the genomic DNAs are cohybridized to oligonucleotide probes attached to an array surface. Hybridized arrays are then scanned to detect the relative DNA quantities at each probe location, giving a ratio measurement between test and reference signal [3, 13]. Measurements are reported as log2 ratios. A schematic of the experimental set-up is given in Figure 2.

In this work, we demonstrate the functional data methods using the publicly available dataset GSE39281 [17], available from the Gene Expression Omnibus database [4]. These data consist of Agilent-022060 SurePrint G3 Human CGH Microarray 4x180K measurements on 93 patients with metastatic urothelial carcinoma, and include 78 patients with MIBC, and 15 patients with non-MIBC.

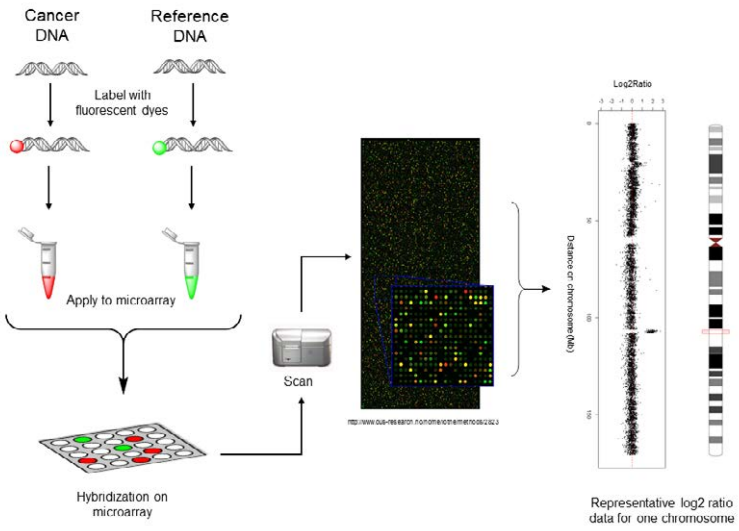

Figure 2: Schematic depiction of the array CGH method. The depiction of the log2 ratio data on the right is a representative illustration from one chromosome from one patient in the dataset used in this work (GSE39281 [17]).

1.2 Main motivation for fitting via functional data methods

Copy number measurements obtained via the measurement modalities described above give information about regional deviations from the normal diploid genomic condition. The basic idea in many CNA studies is to measure the level of genomic loci along the entire chromosome (and for all chromosomes), using a high-density method for assessing the DNA copy number such as hybridization-based SNP arrays and CGH arrays, as well as next-generation sequencing modalities that provide read counts for genomic loci. These measurements can be linearly displayed on chromosome maps, and as such when plotted versus chromosome position, form a (noisy) profile of the DNA content across the chromosome. Deviations from the 'normal' diploid condition (representing gains or losses of regions of genetic material) can be detected as shifts in these profiles. Segments of the chromosome that preserve the 'normal' condition should have consistent mean levels of CN representing the diploid condition; altered regions will have mean levels of CN along their segment that differs from normal. An example of a full genome scan for one patient in the BLCA dataset is given in the upper portion of Figure 3, showing the centering of much of the profile around zero (ie where ratio between tumor and normal reference is equal to one), with CNAs visible as regions of varying width of displacement from the 'normal' condition. A focus on just one chromosome is given in the bottom portion of Figure 3.

This type of representation naturally lends itself to functional data methods, in which the DNA copy number profiles are viewed as functions along the chromosome length, and segmentation methods can be used to define the breakpoints where copy number shifts occur. Functional data methods can be employed to characterize the genomic profiles and explore them via functional data regression methods.

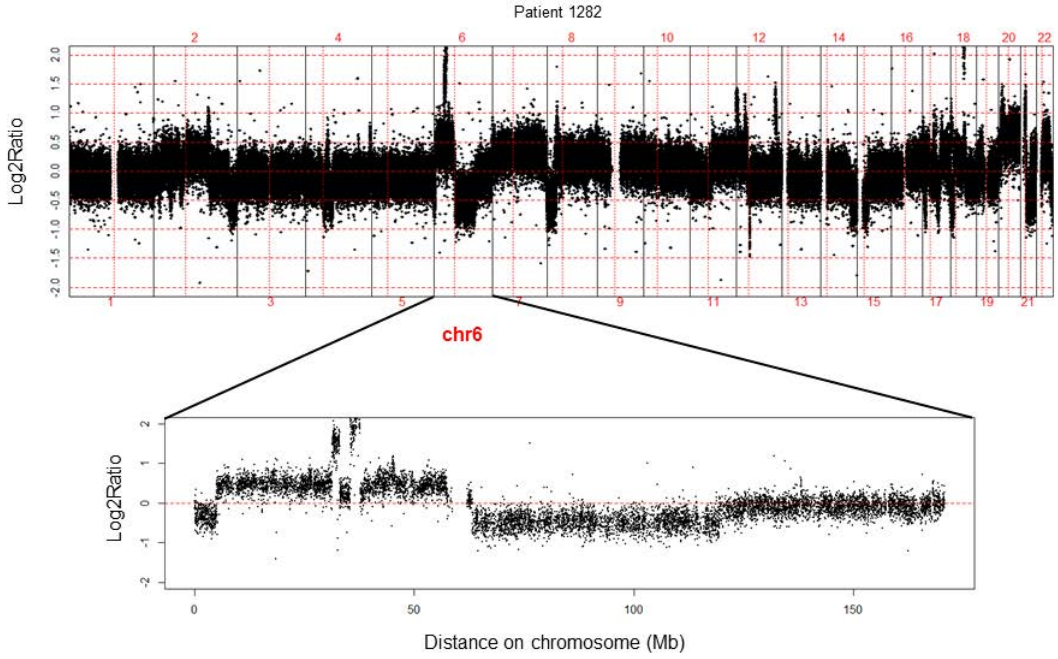

Figure 3: Top: Array CGH scan (log2 ratios) for one patient across 22 chromosomes. Bottom: Zoom in on chromosome six for the representative patient.

2. Wavelet-based Methods for Fitting Copy Number Data

A primary focus of CN studies has been to carry out some form of segmentation of CN data in order to locate the abrupt shifts in the genomic profile, and a variety of methods (including functional data methods) have been proposed to pinpoint and catalog CNAs (reviewed in [12, 19]). These methods have included wavelet-based curve estimation, proposed for CN data by Hsu in 2005 [8], mainly as a method for denoising CN data for downstream analyses. Much less work has been done examining CN data via wavelet basis functions in a Bayesian framework, or employing the machinery of functional regression to explore covariate effects on the genomic profiles viewed as functional outcomes. Wavelets have proven useful for nonparametric function estimation problems because they are capable of handling functions with discontinuities and other inhomogeneities, as is characteristic of CN data in which CNAs appear as 'breaks' in the continuous genomic profile. In addition, wavelet methods often yield sparse representations of functions. Piecewise smooth functions (such as the piecewise constant model for CN data) are well represented sparsely, with larger coefficients associated only with the disjunctions in the functional profile. A brief review of wavelets is given in the following subsection.

2.1 Brief introduction to wavelets

It is typical in nonparametric function estimation to carry out curve representation via some form of basis expansion, often employing smooth basis functions such as splines. Wavelets are a class of functions that possess oscillatory behavior and compact support (or have the property of rapidly decaying to zero if not compactly supported), and allow multiscale representation of underlying functions. Curve fitting using wavelets estimate the unknown profile function with a linear combination of wavelet basis functions, requiring estimation of the regression coefficients from the linear combination. So, for an unknown function $f(x)$, we can decompose it as:

$$f(x) = \beta_0 + \sum_{j=1}^{J-1} \sum_{i=1}^{n(j)} \beta_{ji} B_{ji}(x),$$

with the double summation characterizing the multiple resolution levels of the wavelet bases, and B_{ji} generically representing the wavelet basis functions.

Given a mother wavelet function, *wavelets* are generated via dilation and translation of the mother wavelet functions. The wavelet basis set employed in this work for CN data is the Haar basis, with the mother wavelet function given by:

$$\psi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}) \\ -1 & x \in [\frac{1}{2}, 1) \\ 0 & otherwise \end{cases} \tag{1}$$

For given integers *j*, *k*, the functions formed by:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k) \tag{2}$$

form an orthonormal set; in fact, $\{\psi_{j,k}(x)\}_{j,k \in \mathbb{Z}}$ can be a complete orthonormal basis for $L^2(\mathbb{R})$.

A function *f(x)* $\in$ $L^2(\mathbb{R})$ can be represented into the following generalized Fourier series:

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(x), \tag{3}$$

where $d_{j,k}$ are wavelet coefficients derived according to:

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x) = \langle f, \psi_{j,k} \rangle .$$

The Haar father wavelet (scaling function) is given by:

$$\phi(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & otherwise \end{cases},$$

with finest-level (scale 2^J) father wavelet coefficient given by

$$c_{J,k} = \int_0^1 f(x)2^{J/2}\phi(2^J x - k)dx = \int_0^1 f(x)\phi_{J,k}(x)dx.$$

The associated *Haar father wavelet*s are given by:

$$\phi_{J,k}(x) = \begin{cases} 2^{J/2} & x \in [2^{-J}k, 2^{-J}(k+1)] \\ 0 & otherwise \end{cases}$$

The set of coefficients $\{c_{J,k}\}_{k=0}^{2^J-1}$ and the associated Haar father wavelets at scale J give an *approximation* to a function *f(x)* given by:

$$f_J(x) = \sum_{k=0}^{2^J-1} c_{J,k}\phi_{J,k}(x).$$

For the approximation to a function at adjacent scale levels, the finer approximation at level J is equal to the courser approximation given at level $J-1$, *plus* the additional detail encapsulated in the detail coefficients at the coarser level; so,

$$\begin{aligned} f_{j+1}(x) &= f_j(x) + \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(x) \\ &= \sum_{k=0}^{2^j-1} c_{j,k}\phi_{j,k}(x) + \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(x). \end{aligned}$$

So the Haar father wavelet approximation at finer scale $j+1$ is equivalent to the father wavelet approximation at scale j, plus the details stored in the coefficients $\{d_{j,k}\}_{k=0}^{2^j-1}$. These ideas characterize the multiresolution nature of wavelets, with representations at progressively finer levels of detail, enabling capture of information both in a frequency domain and a location domain.

This characterization effectively means that a general function *f(x)* can be represented as a sum of a 'smooth' or 'kernel-like' part involving the father wavelet $\phi_{j_0,k}$ and a set of detail representations involving the mother wavelet $\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x)$. The $\phi_{j_0,k}$ represents the 'average' or 'overall' level of the function, and the rest represents detail of the function.

Wavelet coefficients represent differences in averages that can be used to represent mean copy number of adjacent chromosomal segments, so are well suited to determining the boundary locations of shifts in copy number.

2.2 Wavelet shrinkage

In wavelet shrinkage, a function is observed contaminated with additive noise. A vector of observations $y = (y_1, \ldots, y_n)$ of a function are assumed to arise from the following model:

$$y_i = g(x_i) + \epsilon_i, \tag{4}$$

for unknown 'true' function *g(x)*. The function is transformed to wavelet domain via a discrete wavelet transform (DWT), where the noisy function's wavelet coefficients undergo *shrinkage* or thresholding. Note that the DWT can be characterized as an orthogonal matrix multiplication, but actual computational implementation proceeds via much faster and efficient algorithms that do not employ matrix manipulations. The key ideas underpinning wavelet shrinkage were introduced in [7]. Subsequently, an inverse wavelet transformation allows estimation of the function. It is typically assumed that observation of the function occurs at equal intervals, and that the $\epsilon_i \sim N(0, \sigma^2)$ and are independent. As hybridization-based CN data is not usually measured at equally spaced genomic loci, application of wavelet methods to CN data require some adjustment to accomodate the uneven measurement grid (see [8] and references therein and [15]).

The DWT is applied to model (4) above, moving from the data space to the wavelet space. Let *W* represent the matrix of the transform, and let *y*, *g*, and *e* be vectors of observations, the true unknown function, and noise, respectively. Then let d^* = *Wy*, *d* = *Wg*, and ϵ = *We*. Then we have the wavelet-transformed model:

$$d^* = d + \epsilon. \tag{5}$$

For functions that are smooth or smooth with discontinuities, the vector *d* is sparse. Since W is orthogonal, the wavelet transform ϵ of the white noise error is also white noise. The main idea of shrinkage is that large values of coefficients in the wavelet domain most likely contain signal and noise, where smaller coefficients are mostly just noise. So thresholding these coefficients forms an estimate $\hat{d}$, by removing coefficients in d^* below the threshold.

2.3 Bayesian wavelet shrinkage

In a Bayesian framework, knowledge that the vector of wavelet coefficients form a sparse set can be leveraged into specification of a prior distribution on the 'true' wavelet coefficients, $d_{j,k}$. Following transforming observed genomic profile data from data domain to wavelet domain via wavelet transform, the posterior distribuition of the wavelet coefficients $d^*_{j,k}$ can be computed using the sparsity prior. Then the inverse wavelet transform is used on the wavelet coefficients posterior distribution (or some value like posterior mean or median of coefficients), to get a Bayesian estimate of the 'true' function.

As in the frequentist setting, denoising of functions or smoothing in the wavelet domain in a Bayesian framework proceeds via some version of thresholding of wavelet coefficients after transforming from data space to wavelet space. This involves prior specification on the wavelet coefficients to enforce sparsity assumption. Several sparsity-inducing priors have been proposed, including a mixture of Gaussians [6], which sets certain coefficients to be very small (but not identically zero). A function that has a truly sparse wavelet transform (such as one that is piecewise smooth with jump discontinuities) can be handled with a sparsity inducing prior that allows for exact zero coefficients to be produced. To achieve this, Abromowich (1998) proposed a mixture of a Gaussian distribution with degenerate point mass at zero to allow some coeffiencts to be forced fully to zero [1]. This idea was extended to a mixture of point mass at zero with a heavy-tailed distribution by Johnstone and Silverman [10], to allow for some true zero coefficients, as well as some (much) larger

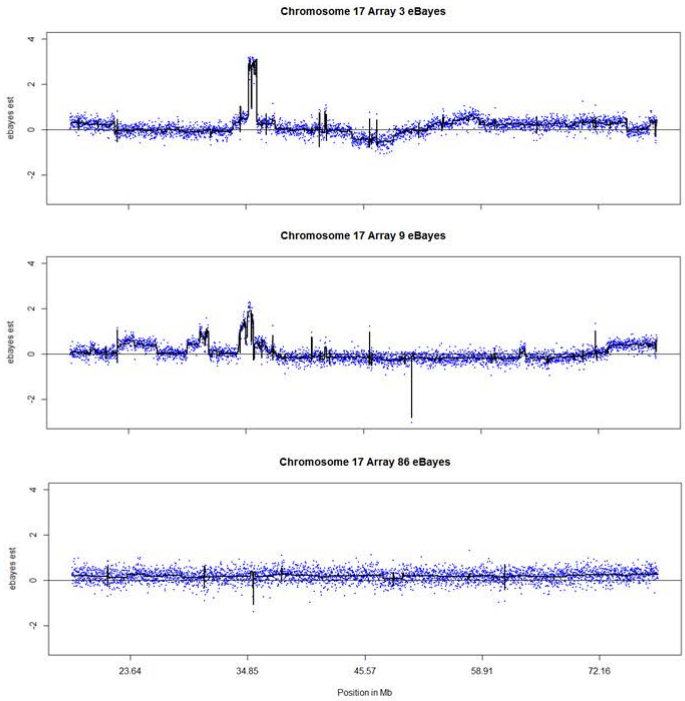

Figure 4: Examples of fitted profiles for chromosome 17 for three patients (solid black line indicates wavelet-based fitted curve).

coefficient values. Empirical Bayes estimates are often used to determine hyperparameters for carrying out shrinkage.

The Johnstone and Silverman prior for wavelet coefficient d uses mixture of heavy-tailed distribution and point mass:

$$\pi(d) = \omega\tau(d) + (1-\omega)\delta_0(d),$$

for mixing weight ω, and τ a heavy tailed distribution that is symmetric, unimodal, and with heaviness not greater than Cauchy distribution.

Mixing weights are estimated by marginal maximum likelihood then plugged in to prior to produce posterior distribution of the coefficients.

Examples of fitted profiles for several patients in the GSE39281 dataset are given in Figure 4. In this figure, chromosome 17 is shown from three patients, with wavelet-based fitted curves for each profile overlaid in the solid black line. Datapoints represent log2 ratio measurements of tumor versus normal probe reading for that probe locus, and both 'wider' events of CNA, as well as the ability of the estimation procedures to capture narrow, focal events, are apparent.

3. Wavelet-based Functional Data Regression for Copy Number Data

Functional data regression methods have received much recent attention and methodological development, as they allow for characterizing covariate effects on functional outcomes, treating the entire function as the object of analysis, or allowing for covariates themselves that are functional in nature (see [14] for a recent review). The methods received huge impetus via their discussion in one of the key resources on functional data analysis, Ramsay and Silverman's 2006 monograph on the topic ([16].

In the present work, we examine functional response regression for CNA data, viewing the genomic copy number profiles as functional outcomes, and attempting to assess whether mean CN profiles differ between MIBC and non-MIBC patients. In bladder cancer, CNAs are thought to be more frequently associated with MIBC as opposed to non-MIBC cases [11]. Some alterations, such as alterations in Chromosome 9, can be quite common in BLCA (>50% prevalence, in both MIBC and non-MIBC patients), while other alterations can be more common in one group compared to another [11]. Delineating what patient characteristics (including muscle-invasiveness) are associated with CNA patterns can lead to important clues into the mechanistic process of cancer progression and large-scale genomic instability, and functional data regression methods can provide a valuable tool in identifying possible drivers of progression pathways (see discussion in [11]).

The general question in this case is whether the shape of the overall profile (function) depends on to which categorical class the profile belongs. For the CN data, for instance, each patient has a copy number profile along the genome (or along a given chromosome), and it might be of interest whether those profiles differ between categorical groups of patients (eg treated versus untreated, MIBC versus non-MIBC, etc).

The general model is given by:

$$f_{ik}(t) = \mu(t) + \alpha_k(t) + \epsilon_{ik}(t),$$

where $f_{ik}(t)$ is the profile for the i^{th} individual in the k^{th} categorical group, $\mu(t)$ is the overall mean response across all individuals, and the $\alpha_k(t)$ functions are the effect functions for each categorical group, representing departures from the overall profile that are group-specific. The residual functions $\epsilon_{ik}(t)$ capture residual variation left over after explaining the outcome function using the information contained in the group categories. For scalar predictors, the design matrix X containing values of the p predictor variables uses the scalar values of the predictors rather than the *(0, 1)* coding used for categorical predictors.

The inferential goal here is to estimate the functional parameters μ and α_k using the data from the profiles, and to determine if differences exist between profile groups based on group membership.

In matrix notation, with information on the p predictors in design matrix *X*, with observations of function $f_{ik}(t)$ given by vector $y_i(t_j)$ for individual i at time j, $j = 1, \ldots, N_i$, we have:

$$Y_i(t_j) = \sum_{a=1}^{p} X_{ia} B_a(t_j) + E_i(t_j), \tag{6}$$

where coefficient $B_a(t)$ is the effect of predictor X_a on the functional response at t.

Functional data regression using wavelet bases is especially challenging given the inhomogeneous nature of the profiles. Ideally, the goal is to be able to examine the CN profiles, where individuals have discontinuities at different locations, over different portions of the support space, and of differing amplitude, and be able to connect the patterns of response to underlying clinical features that define the patient groups.

A key limitation to carrying out wavelet based function-on-scalar regression models is the availability of software to implement the procedures described above. We have carried out analyses of the public dataset GSE39281 using the WFMM software made available by Morris, *et al*, and described in [15]. This software implements Bayesian estimation of the functional wavelet mixed effects models for replicated functional data, and can be modified to handle a variety of regression models. Implementation was in Matlab2015a.

3.1 Application to urothelial bladder carcinoma patient data

Functional response regression methods allow us to use the information in genomic profiles of multiple individuals who have either muscle invasive or non-muscle invasive bladder carcinoma. A goal of the analysis is characterization of whether overall mean genomic response profiles differ between the groups, either in terms of location of key features or their amplitude.

Results for model fitting procedures for chromosome 6 appear in Figure 5, with non-MIBC mean profile shown in red and MIBC profile shown in blue. As is apparent, there is substantial overlap between the profiles for the different groups, including where they are deviating from the normal diploid condition, indicating that, for this chromosome at least, the nature of large scale structural genomic alterations do not strongly differentiate muscle invasive from non-muscle invasive disease. It is interesting to note that there are some noticable differences in amplitude for certain focal events, however, which are noted in Figure 5. Intriguingly, several of these highlighted narrow scale events are associated with genes known from other types of studies to be associated specifically with bladder cancer or that are connected to immune response, which is of particular interest in our study.

3.2 Ongoing work

Work using functional data regression methods to fully characterize these structural genomic events, differentiating MIBC and non-MIBC patients, continues to proceed, with special emphasis on developing formal inference. We are also continuing our development of novel data-adaptive shrinkage prior specifications, geared towards incorporating prior knowledge specific to the copy number setting. In particular, we are developing novel priors for use in wavelet shrinkage that allow capturing of information about event width. Inferential procedures employed with CN profiles have focused on event amplitude and overlap between individiuals, but have failed to address event width (and the concommitent issues of non-independence of genes within broad regions). This is an exciting area to be applying functional data regression methods and they provide an important tool capable of addressing key questions in copy number biology that have been inaccesible to previous methods.

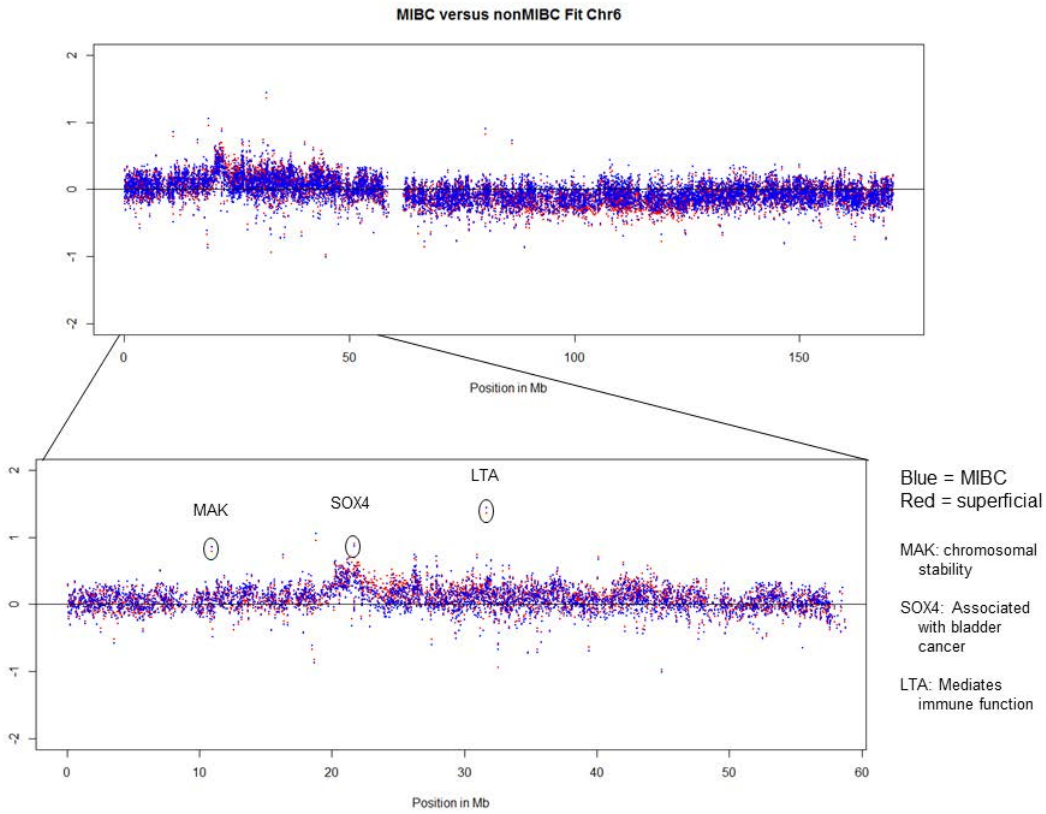

Figure 5: Wavelet regression fits of MIBC (blue) and non-MIBC (red) BLCA data for Chromosome 6.

References

[1] Felix Abramovich, Theofanis Sapatinas, and Bernard W Silverman. Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749, 1998.

[2] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34(4):369–376, 2003.

[3] Donna G Albertson and Daniel Pinkel. Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics*, 12(suppl 2):R145–R152, 2003.

[4] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[5] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010.

[6] Hugh A Chipman, Eric D Kolaczyk, and Robert E McCulloch. Adaptive bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92(440):1413–1421, 1997.

[7] David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[8] Li Hsu, Steven G Self, Douglas Grove, Tim Randolph, Kai Wang, Jeffrey J Delrow, Lenora Loo, and Peggy Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, 2005.

[9] Norman Huang, Parantu K Shah, and Cheng Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13(3):305–316, 2012.

[10] Iain M Johnstone and Bernard W Silverman. Empirical bayes selection of wavelet thresholds. *Annals of Statistics*, pages 1700–1752, 2005.

[11] Margaret A Knowles and Carolyn D Hurst. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, 15(1):25–41, 2015.

[12] Weil R Lai, Mark D Johnson, Raju Kucherlapati, and Peter J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21(19):3763–3770, 2005.

[13] Evi Michels, Katleen De Preter, Nadine Van Roy, and Frank Speleman. Detection of dna copy number alterations in cancer by array comparative genomic hybridization. *Genetics in Medicine*, 9(9):574–584, 2007.

[14] Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2(1):321–359, 2015.

[15] Jeffrey S Morris and Raymond J Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199, 2006.

[16] James O Ramsay. *Functional data analysis, 2nd ed*. Wiley Online Library, 2006.

[17] Markus Riester, Lillian Werner, Joaquim Bellmunt, Shamini Selvarajah, Elizabeth A Guancial, Barbara A Weir, Edward C Stack, Rachel S Park, Robert O'Brien, Fabio AB Schutz, et al. Integrative analysis of 1q23. 3 copy-number gain in metastatic urothelial carcinoma. *Clinical Cancer Research*, 20(7):1873–1883, 2014.

[18] Louise V Wain, John AL Armour, and Martin D Tobin. Genomic copy number variation, human health, and disease. *The Lancet*, 374(9686):340–350, 2009.

[19] Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

[20] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10:451–481, 2009.