Cristiano *et al. BMC Cancer* (2020) 20:856
https://doi.org/10.1186/s12885-020-07304-3

BMC Cancer

RESEARCH ARTICLE Open Access

# Bayesian copy number detection and association in large-scale studies

Stephen Cristiano[1†], David McKean[2†], Jacob Carey[1†], Paige Bracci[3], Paul Brennan[4], Michael Chou[5], Mengmeng Du[6], Steven Gallinger[7], Michael G. Goggins[8,9], Manal M. Hassan[10], Rayjean J. Hung[7], Robert C. Kurtz[11], Donghui Li[12], Lingeng Lu[13], Rachel Neale[14], Sara Olson[6], Gloria Petersen[15], Kari G. Rabe[15], Jack Fu[1], Harvey Risch[13], Gary L. Rosner[1,10], Ingo Ruczinski[1], Alison P. Klein[2,5,9*] and Robert B. Scharpf[1,2*]

## Abstract

**Background:** Germline copy number variants (CNVs) increase risk for many diseases, yet detection of CNVs and quantifying their contribution to disease risk in large-scale studies is challenging due to biological and technical sources of heterogeneity that vary across the genome within and between samples.

**Methods:** We developed an approach called CNPBayes to identify latent batch effects in genome-wide association studies involving copy number, to provide probabilistic estimates of integer copy number across the estimated batches, and to fully integrate the copy number uncertainty in the association model for disease.

**Results:** Applying a hidden Markov model (HMM) to identify CNVs in a large multi-site Pancreatic Cancer Case Control study (PanC4) of 7598 participants, we found CNV inference was highly sensitive to technical noise that varied appreciably among participants. Applying CNPBayes to this dataset, we found that the major sources of technical variation were linked to sample processing by the centralized laboratory and not the individual study sites. Modeling the latent batch effects at each CNV region hierarchically, we developed probabilistic estimates of copy number that were directly incorporated in a Bayesian regression model for pancreatic cancer risk. Candidate associations aided by this approach include deletions of 8q24 near regulatory elements of the tumor oncogene *MYC* and of Tumor Suppressor Candidate 3 (*TUSC3*).

**Conclusions:** Laboratory effects may not account for the major sources of technical variation in genome-wide association studies. This study provides a robust Bayesian inferential framework for identifying latent batch effects, estimating copy number, and evaluating the role of copy number in heritable diseases.

**Keywords:** Pancreatic cancer, SNP array, Copy number variants, Genome-wide association, CNPBayes, Batch effects

*Correspondence: aklein1@jhmi.edu; rscharpf@jhu.edu
†Stephen Cristiano, David McKean, and Jacob Carey contributed equally to this work.
[2]Department of Oncology The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Full list of author information is available at the end of the article

## Background

Germline copy number variants (CNVs) can be identified from hybridization-based arrays and capture-based sequencing with measures of abundance derived from intensities and normalized read depth, respectively. Biological and technical sources of heterogeneity of these measurements are intricately related. For example, the GC composition of genomic DNA effects PCR efficiency and leads to autocorrelated measures of DNA abundance across the genome [1–4]. These effects have been shown to be heterogeneous across the genome and to differ in both magnitude and direction between samples [1, 5, 6]. Hidden Markov models and nonparametric segmentation algorithms for CNV detection over-segment low-quality data where these effects are the most pronounced, contributing to false positive deletion and duplication calls.

For studies with hundreds to thousands of samples, estimation of copy number at regions known to harbor CNVs has the potential to improve sensitivity and specificity as technical sources of variation across the genome are largely controlled when limited to a focal genomic region (less than 1 MB) and variation between samples can be explicitly modeled [7–12]. Such CNV regions are of particular interest for a comprehensive assessment of common genetic variants and their relationship to disease. However, scaling marginal models to CNV regions across the genome and to thousands of samples has proved challenging. The sources of technical variation giving rise to batch effects are typically unknown. Standard approaches for estimating latent batch effects in high-throughput experiments such as surrogate variable analysis are not appropriate when the biological variation of interest (copy number) is also unknown [13]. In addition, the statistical framework for copy number estimation must flexibly accommodate deletions and duplications of varying size and allele frequencies. Symptomatic of the challenges in copy number analyses and the limitations of current methods, genome-wide association studies rarely incorporate copy number in the initial publication despite their well established role in neurodevelopmental disorders [14–16] and cancer [17]. Previous genome-wide studies of pancreatic cancer and copy number have been limited in size with fewer than 250 pancreatic cancer patients [18, 19].

Here, we performed genome-wide copy number analysis for 3,974 cases and 3,624 controls in PanC4 using Illumina's OmniExpress Exome array. We developed methods for identifying latent batch effects at CNV regions from commonly available experimental data on the samples. The effects of copy number and batch on measures of DNA abundance were modeled hierarchically through implementation of Bayesian finite mixture models. For the association model, we used Markov Chain Monte Carlo (MCMC) to incorporate the uncertainty of the integer copy numbers in a logistic regression model of pancreatic cancer risk.

## Methods

*The Pancreatic Cancer Case and Control Consortium*

Clinical and demographic characteristics of the cases and controls in PanC4 and recruitment methods have been previously described [20]. All samples were processed using GenomeStudio (version 2011.1, Genotyping Module 1.9.4). For GC-correction, we sampled a random subset of 30,000 Illumina probes, fit LOESS with span 1/3 to the scatterplot of $\log_2 R$ ratios and probe GC content, and predicted the $\log_2 R$ ratios for the full probeset from the LOESS model. For spatial correction, we applied LOESS to the GC-corrected $\log_2 R$ ratios at single nucleotide polymorphisms (SNPs) with balanced allele frequencies ($0.4 <$ B allele frequency $< 0.6$) ordered by genomic position within each chromosome arm and predicted the GC-corrected $\log_2 R$ ratios for the full probeset, including SNPs with imbalanced allele frequencies. The residuals from the spatial LOESS were used in all downstream analyses with CNPBayes.

*CNV regions:*

CNV regions identified for further analysis by CNPBayes were obtained from the collection of CNVs identified from a hidden Markov model as well as known CNV regions from the 1000 Genomes Project. For the former, we fit a 5-state hidden Markov model implemented in the R package VanillaICE (version 1.40.0) using default parameter settings [21]. To obtain a high confidence call set, we removed CNVs with fewer than 10 probes, CNVs with posterior probability less than 0.9, and restricted inference to autosomal chromosomes. To assess the effect of spatial adjustment on copy number inference, we stratified the samples into deciles of median absolute deviation and autocorrelation coefficient (ACF) and compared the results of the 5-state HMM fit after GC-correction to the CNVs identified after spatial correction. Concordance of CNVs identified by the HMMs was defined by $\geq 50\%$ reciprocal overlap [22].

CNV regions were defined by the set of non-overlapping disjoint intervals across the pooled set of all CNVs from cases and controls. We computed the number of subjects with a CNV overlapping each disjoint interval, retaining intervals where CNVs were identified in at least 150 participants. Among the disjoint intervals, we defined the CNV region as the minimum start and maximum end for which at least 50 percent of the copy number altered samples had a CNV. For CNV regions obtained from the 1000 Genomes Project, we excluded regions that did not span at least 4 markers on the OmniExpress array.

*Batch effects:*

We evaluated both chemistry plate and DNA extraction method as surrogates for batch effects. With provisionally

defined batches by plate or extraction method, we compared the empirical cummulative distribution function (eCDF) of the mean $\log_2 R$ ratio between two batches (excluding samples with $\log_2 R$ ratio $< -1$) by the Kolmogorov-Smirnov (K-S) test statistic. For two batches without a statistically significant difference in the K-S statistic at a type 1 error of 0.01, the batches were combined into a single new batch. This procedure was applied recursively at each CNV region until no further grouping of batch surrogates could be obtained.

*Hierarchical Bayesian mixture model:*

Hierarchical Bayesian mixtures of $t$-distributions were used to cluster median $\log_2 R$ ratios within a CNV region. Let $r_{ib}$ and $z_{ib}$ denote the observed one-dimensional summary of $\log_2$ ratios measured from the array and the true (but latent) mixture component, respectively, for the $i$th individual in batch $b$. Given $z_{ib}$ is some integer $h$ ($h \in \{1, \ldots, K\}$ for a $K$-component model), our sampling model for the observed data is a shifted and scaled $t$-distribution with $d$ degrees of freedom, mean $\theta_{hb}$, and standard deviation $\sigma_{hb}$ that depends on batch:

$$\begin{aligned}
\left[r_{ib}|z_{ib}=h,\theta_{hb},\sigma^2_{hb},U_{ib}\right] &\sim \text{Normal}\left(\theta_{hb}, \frac{\sigma^2_{hb}}{U_{ib}/d}\right),\\
z_{ib}|\pi_{b1},\ldots,\pi_{bK} &\sim \text{Multinomial}(\pi_{b1},\ldots,\pi_{bK}),\\
\pi_{b1},\ldots,\pi_{bK}|\alpha^\pi_1,\ldots,\alpha^\pi_K &\sim \text{Dirichlet}\left(\alpha^\pi_1,\ldots,\alpha^\pi_K\right), \text{ and}\\
U_{ib}|d &\sim \text{Gamma}(d/2, d/2).
\end{aligned}$$

The degrees of freedom $d$ controls robustness to outliers with larger values approximating a mixture of normal distributions [23]. To stabilize the mean and precision of batches having fewer samples, we model these parameters hierarchically with computationally convenient conjugate priors. Our sampling model for the batch means is normal and the precisions are Gamma,

$$\begin{aligned}
\theta_{hb}|\mu_h,\tau^2_h &\sim \text{Normal}\left(\mu_h,\tau^2_h\right) \text{ and}\\
\tilde{\sigma}^2_{hb}|\nu_0,\sigma^2_0 &\sim \text{Gamma}\left(\frac{1}{2}\nu_0, \frac{1}{2}\nu_0\sigma^2_0\right),
\end{aligned}$$

with $\mu_h$ representing the overall mean for component $h$, $\tau_h$ capturing the heterogeneity of the batch-specific means, and $\tilde{\sigma}^2_{hb} = 1/\sigma^2_{hb}$. Conjugate priors on $\mu_h$, $\tau^2_h$, $\sigma_0$, and $\nu_0$ are given by

$$\begin{aligned}
\mu_h|\mu_o,\tau^2_0 &\sim \text{Normal}(\mu_0,\tau^2_0) \text{ for } h=1,\ldots,K,\\
\tilde{\tau}^2_h &\sim \text{Gamma}\left(\frac{1}{2}\eta_0, \frac{1}{2}\eta_0 m^2_0\right),\\
\sigma^2_0|a_0,b_0 &\sim \text{Gamma}(a_0,b_0), \text{ and}\\
\nu_0|\beta &\sim \text{Geometric}(\beta).
\end{aligned}$$

Label switching is well known in Bayesian mixture models. In addition to visual inspection of the chains, we compared the ordering of parameter means for subsequences of the chains. Label switching occurred most often when the number of components specified was too large and these models were discarded. In addition, we use an informative prior on $\tau^2_h$ that governs the heterogeneity of the mean for mixture component $h$ across the batches (Table S3). This prior discourages label switching at bona fide copy number polymorphisms since this would typically result in a large variance of the component means.

As all priors were conjugate, we used Gibbs sampling to approximate the joint posterior distribution of $p\left(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta},\boldsymbol{\sigma}^2,\boldsymbol{z},\boldsymbol{\pi},\nu_0,\tau^2_0,\sigma^2_0,m^2_0,\eta_0|\boldsymbol{r},K,d\right)$. We refer to the above implementation of the Gibbs sampler with batch-specific means and variances as the multi-batch (MB) model. CNPBayes provides several more parsimonious alternatives to the MB model, including a pooled variance model (MBP) with a single variance estimate per batch. In addition, we evaluated models with a single batch (SB) and a single batch model with pooled variances (SBP) that are special cases of the MB and MBP models, respectively. Hyper-parameters used in the MB, MBP, SB, and SBP models were the defaults in version 1.11.2 of CNPBayes (Table S3).

*Implementation:*

Heavy-tailed marginal distributions of the one-dimensional $\log_2 R$ ratio summaries were often a consequence of batch effects. When latent batch effects were estimated as previously described, near-Gaussian mixtures were needed to fit the very peaked densities of log ratios near the central mode. As residual outliers and lack of normality after taking batch effects into account were often asymmetric and could be captured by an additional mixture component, we fit finite mixtures of near-Gaussian distributions with $d$=100 degrees of freedom in both the PanC4 application and simulations. Estimation of $d$, for example from a discrete uniform prior ([24, 25]), is not currently available in CNPBayes.

For studies of germline CNVs, extreme observations in the left-tail typically correspond to homozygous deletions and, when rare, may be present in a subset of the estimated batches. The consequences of a rare deletion present in a subset of batches are two-fold: (1) due to the hierarchical nature of the model, a mixture-component with a very large variance will be needed to accommodate the extreme observations and (2) the mixture component indices may correspond to different copy number states between batches, complicating subsequent efforts to map mixture component indices to integer copy numbers. Rather than exclude these observations, we augment the observed data with simulated homozygous deletions. The simulated observations ensure the mixture component indices capture the same latent copy number in each batch. We rationalize this approach as being comparable to an empirically derived prior that large negative values at such germline CNV regions are

not outliers of the hemizygous and diploid states but bona fide homozygous deletions. Since our model does not assume a one-to-one mapping between mixture components and copy number nor that any of the alterations identified will be in Hardy Weinberg equilibrium (HWE), the assessment of HWE for germline CNVs can be a useful post-hoc quality control. While evidence against HWE does not necessarily indicate problems with the CNV calling, support for HWE would be unlikely if there were major sources of technical variation not yet accounted for.

As fitting hierarchical Bayesian mixture models is computationally intensive, we implemented ad hoc procedures to reduce computation (see also Scalability and Software). First, we considered only 3 and 4 component models when homozygous deletions were apparent (one- and two-component models were not evaluated). Secondly, MB and MBP models were only evaluated when more than 2% of the samples had a posterior probability $< 0.99$ in the SB and/or SBP models. Thirdly, for each model under consideration, we independently initialized 10 models with parameters randomly sampled from their priors and ran a short burnin of 200 iterations for each model. Only the model with the largest log likelihood was selected for an additional 500 burnin simulations and 1000 simulations post-burnin. Finally, to aid comparison between hierarchical SB, SBP, MB, and MBP models, the CNPBayes package implements Chib's method to estimate the marginal likelihood [26], allowing estimates of the relative evidence between two models by Bayes factors. However, as estimation of the marginal likelihoods requires additional MCMC simulations, we only computed marginal likelihoods when the difference of simple post-hoc statistical summaries, such as the log likelihood evaluated at the last iteration, was small (e.g., $< 10$). CNPBayes automatically provides posterior predictive distributions of the CNV region summaries for goodness of fit assessments, allowing simple verification that the selected model is not simply the best of many poor fitting models. We recommend running multiple chains to assess convergence [27] and additional MCMC simulations with an increased thin parameter when autocorrelation is substantial.

*Genotyping mixture components:*

For genotyping the mixture components at a CNP region, our goal was to identify the set of integer copy numbers that would most likely give rise to the observed B allele frequencies (BAFs) at SNPs in this genomic region. We excluded samples that were not assigned to a single mixture component with high posterior probability since these would be less informative. Denoting the mapping of mixture component indices $h$ to integer copy numbers by $f(h)$, the likelihood across SNPs indexed by $j$ and samples indexed by $i$ conditional on the mapping is

$$
\begin{aligned}
p(\boldsymbol{b}|f(\boldsymbol{h}), \boldsymbol{\psi}) &= \prod_i \prod_j p(b_{ij}|f(h_i), \boldsymbol{\psi}), \\
p(b_{ij}|f(h_i), \boldsymbol{\psi}) &= \sum_{g \in G} p(b_{ij}|\boldsymbol{\psi}_g) p(g|f(h_i)), \\
p(b_{ij}|\boldsymbol{\psi}_g) &= \text{dbeta}(b_{ij}|\boldsymbol{\psi}_g), \text{ and} \\
p(g|f(h_i)) &= \text{dbinom}\left(g|p_{jB}, |G(f(h))| - 1\right), \text{ where}
\end{aligned}
$$

dbeta and dbinom are shorthand for the densities for the beta and binomial distributions. For the binomial density, $|\cdot|$ denotes the cardinality of the set and $p_{jB}$ the frequency of the B allele at SNP $j$ in the population of PanC4 participants. The above likelihood averages over the set $G$ of possible allele specific copy numbers ordered by the number of B alleles and indexed by $g$ (e.g., $G(2) \in \{AA, AB, BB\}$; Table S4). Shape and scale parameters ($\boldsymbol{\psi}_g$) for the Beta distribution conditional on the allelic copy numbers are provided in Table S5. Evaluating one-to-one (e.g., $f(\{1,2,3\}) \rightarrow \{0,1,2\}$ for a deletion polymorphism) and many-to-one mappings (e.g., $f(\{1,2,3\}) \rightarrow \{2,2,2\}$), we selected the mapping that maximized the above likelihood on the log-scale.

*Simulation:*

Affymetrix 6.0 data for 990 phase 3 HapMap samples processed on 16 chemistry plates were obtained from Wellcome Sanger Institute (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) [28]. A region on chr4 70,122,981-70,231,746 containing 53 non-polymorphic markers and 1 SNP spans a common copy number polymorphism. To establish a baseline for which both CNPBayes and CNVCALL correctly identify the copy number for all samples, we subtracted 3 from the $\log_2 R$ ratios for samples with apparent homozygous deletions. To simulate batch effects, we simulated a Bernoulli random variable with probability of success 0.5 for each of the 16 chemistry plates. For a plate $k$ where the Bernoulli random deviate was 1, we rescaled the data by a factor $\xi$ and shifted the means by a normal random deviate centered at $\delta$ such that the simulated $\log_2 R$ ratio ($r^*$) for marker $i$ in sample $j$ with integer copy number $c$ becomes $r^*_{ijk} = (r_{ijk} - \bar{r}_c) \times \xi + \bar{r}_c + \epsilon_{ijk}$, where $\epsilon_{ijk} \sim N(\delta, 0.02^2)$ for values of $\delta \in \{0, 0.3, 0.4, 0.5\}$ and $\xi \in \{1, 1.25, 1.50, 1.75, 2\}$. Applying CNVCALL to this data, the matrix of $\boldsymbol{r}^*$ was summarized by the first principal component and mixture models with 3-5 components were evaluated using default parameters. As CNVCALL merges mixture components based on the extent of overlap of the component-specific densities but does not genotype the merged mixture components, we subtracted one from the merged mixture component indices. For CNPBayes, we explored SB, SBP, MB, and MBP models of 3 - 4 components with chemistry plate as the surrogate variable, median $\boldsymbol{r}^*_i$ as one-dimensional summaries for each sample, and default values for hyperparameters. Mixture components were

genotyped using the BAFs from the SNPs in this region as previously described.

*Bayesian logistic regression model for pancreatic cancer:*

For each CNP region, we modeled the case-control status $y_i$ for individual $i$ as:

$$\begin{aligned}
\left[y_i|\boldsymbol{\gamma}, \boldsymbol{X}_i, z, \beta, C_i\right] &\sim \text{Bernoulli}(\theta_i),\\
\text{logit}(\theta_i) &= \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{male}_i + \beta_3 \text{PC1}_i\\
&\quad + \beta_4 \text{PC2}_i + \beta_5 \text{PC3}_i + \beta_6 I_{[\text{high quality}]}\\
&\quad + z \times \left(\beta_7 C_i + \beta_8 C_i * I_{[\text{high quality}]}\right)\\
\beta_j &\sim \text{Cauchy}(0, 2.5^2) \text{ for } j = 0, \ldots, 8,\\
z &\sim \text{Bernoulli}(0.5), \text{ and}\\
C_i &\sim \text{Multinomial}(\pi^*_{i1}, \ldots, \pi^*_{iG}), \text{ where}\\
\pi^*_{ig} &= \sum_{h:h\in\{f(h)=g\}} \pi_{ih}.
\end{aligned}$$

All continuous independent variables were mean centered, including PC1, PC2, and PC3 denoting the first three principal components of the SNP genotype matrix in PanC4 [20]. An indicator for the collection of high quality samples for CNV analyses, $I_{[\text{high quality}]}$, was defined as 1 for samples in this set and 0 otherwise. As the integer copy number $C_i$ was not observed, we treated $C_i$ as a parameter measured with error given by the aggregated posterior probabilities of the mixture component indices after genotyping, $\pi^*_{ig}$. We used JAGS version 4.3.0 with a thin parameter of 25 and 5000 iterations to obtain posterior distributions of these parameters by MCMC [29].

*Scalability and software:*

Hierarchical mixture models were fit to a random sample of 1000 observations from the 7,598 available participants at each CNV region, and to all samples with apparent homozygous deletions. We parallelized our analysis so that all regions were evaluated simultaneously. Bayesian logistic regression models fit independently to each CNV region were also evaluated in parallel. CNPBayes is available from github (https://github.com/scristia/CNPBayes).

## Results

### *Overview of study*

DNA specimens from 7598 European ancestry participants in this consortium were collected at 9 study sites using varying methods of DNA extraction [20]. Randomization of samples to chemistry plates, DNA amplification by PCR, and SNP genotyping using Illumina's OmniExpress Exome-8 array were performed centrally at the Center for Inherited Disease Research (CIDR) (Fig. 1). CNV regions were extracted from the 1000 Genomes project [30] or identified from analysis of the PanC4 samples. Low-level copy number summaries were obtained for each participant by computing the median $\log_2 R$ ratios across available markers from the Illumina array spanned by the candidate CNV region. Independently for each region, we identified latent batch effects in the low-level summaries and fit a Bayesian hierarchical mixture model across the estimated batches using CNPBayes. To model the relationship between copy number and pancreatic cancer risk, we fit a Bayesian logistic regression model that included integer copy number as a covariate measured with error. The copy number measurement error for each participant was obtained from the posterior probabilities in the CNPBayes hierarchical model.

### *Copy number analyses*

$\text{Log}_2 R$ ratios for each participant were GC-corrected using loess. Measures of data quality following GC-correction include the median absolute deviation and lag-10 autocorrelation of autosomal $\log_2 R$ ratios ordered by genomic position. Data quality was high for the majority of PanC4 participants (Figure S1A), though approximately 11% of participants had autocorrelations greater than 0.1. To reduce the spatial autocorrelation, we developed a scatterplot smoother for the $\log_2 R$ ratios that was locally weighted by genomic position (Methods). Following the spatial correction, nearly all samples ($\approx$ 98%) had low autocorrelation (Figure S1B). Rare and common CNVs identified by a 5-state HMM before and after spatial correction revealed near perfect concordance for samples in the first nine deciles of ACF (high quality samples) with sharply lower concordance among samples in the highest decile irrespective of CNV size (Figure S2). Hereafter, we refer to the set of 1,560 samples in the highest decile of ACF (ACF $\geq$ 0.06) as low quality samples and the remaining 6,038 samples in the first nine deciles (ACF $<$ 0.06) as high quality samples.

To evaluate whether copy number inference could be improved by multi-sample methods that directly incorporate batch and other technical sources of variation between samples, we focused our analysis on 217 regions from the 1000 Genomes Project where CNVs were reported in at least 0.1% of European ancestry participants and that encompassed at least four probes on the Illumina OmniExome array (Table S1). Additionally, we identified 46 regions for which deletions or duplications were identified in at least 2% of the PanC4 participants by the HMM applied to the spatially corrected $\log_2 R$ ratios. Collectively, the 263 regions comprised 11.5 Mb of the coding genome and 6.4 Mb of the non-coding genome.

Available multi-sample methods for modeling copy number assume the major sources of batch effects are known (e.g., laboratory or study site). Here, DNA samples were collected from multiple study sites and processed on 94 chemistry plates at a central lab. To identify batch surrogates for the central lab, we developed an approach for grouping chemistry plates with a similar median $\log_2 R$

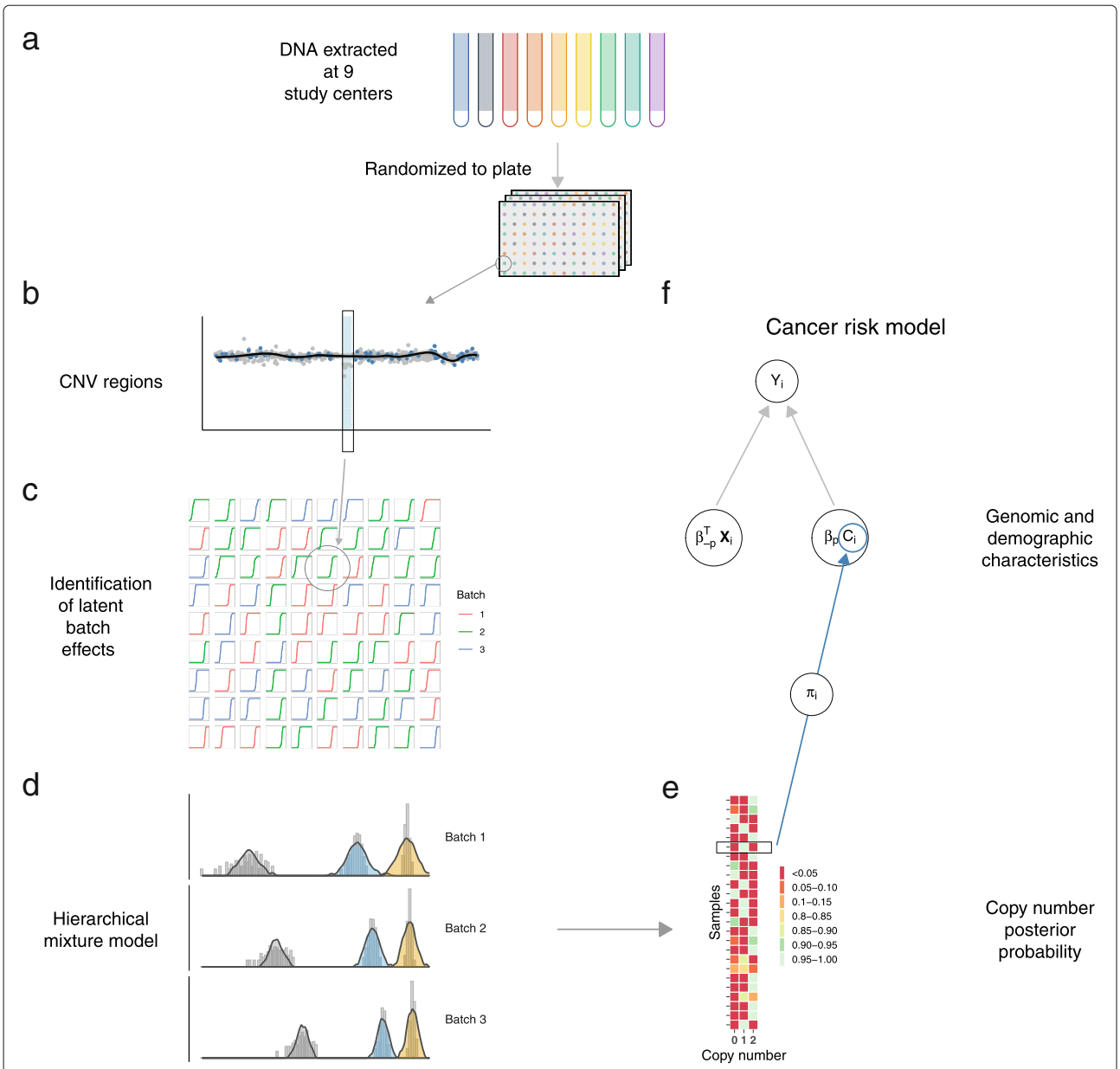


**Fig. 1** Overview of sample processing, estimation of batch effects and copy number, and risk model for pancreatic cancer. **a** DNA samples for pancreatic cancer cases and healthy controls were obtained from 9 different study centers and processed centrally where samples were randomized to chemistry plates. **b** Initial preprocessing of these samples identified candidate CNV regions. As the principal sources of batch effects were unknown, we developed an approach to identify latent batch effects by clustering empirical cummulative distribution functions (eCDFs) of CNV region summaries (**c**) and to genotype these samples via a Bayesian hierarchical mixture model (**d**). Uncertainty of the copy number genotypes (**e**) was propagated from the genomic analyses to the Bayesian logistic regression model for pancreatic cancer risk (**f**)

ratio eCDF (Fig. 2a and b). As an example of these sources of heterogeneity at a single CNV region on chromosome 4, we summarized the $\log_2 R$ ratios for 6,026 high quality samples by the first principal component (PC1) and stratified the PC1 summaries by study site (Fig. 3a) or PCR batch surrogates (Fig. 3b). While the density of PC1 is bimodal when stratified by study site and consistent with a copy number polymorphism, stratification by the eCDF-derived batch surrogates revealed obvious batch effects (e.g., plate group C with 567 samples and plate group E with 786 samples; Fig. 3b). As PCR efficiency is known to be affected by GC content and can vary along the genome, we identified batch surrogates for each CNV region. The median number of batches across the 263 CNV regions was 4 with multiple batches identified for the majority of regions (Figure S3).

Our sampling model for the median $\log_2 R$ ratio is a mixture of $t$ distributions with component-specific means and

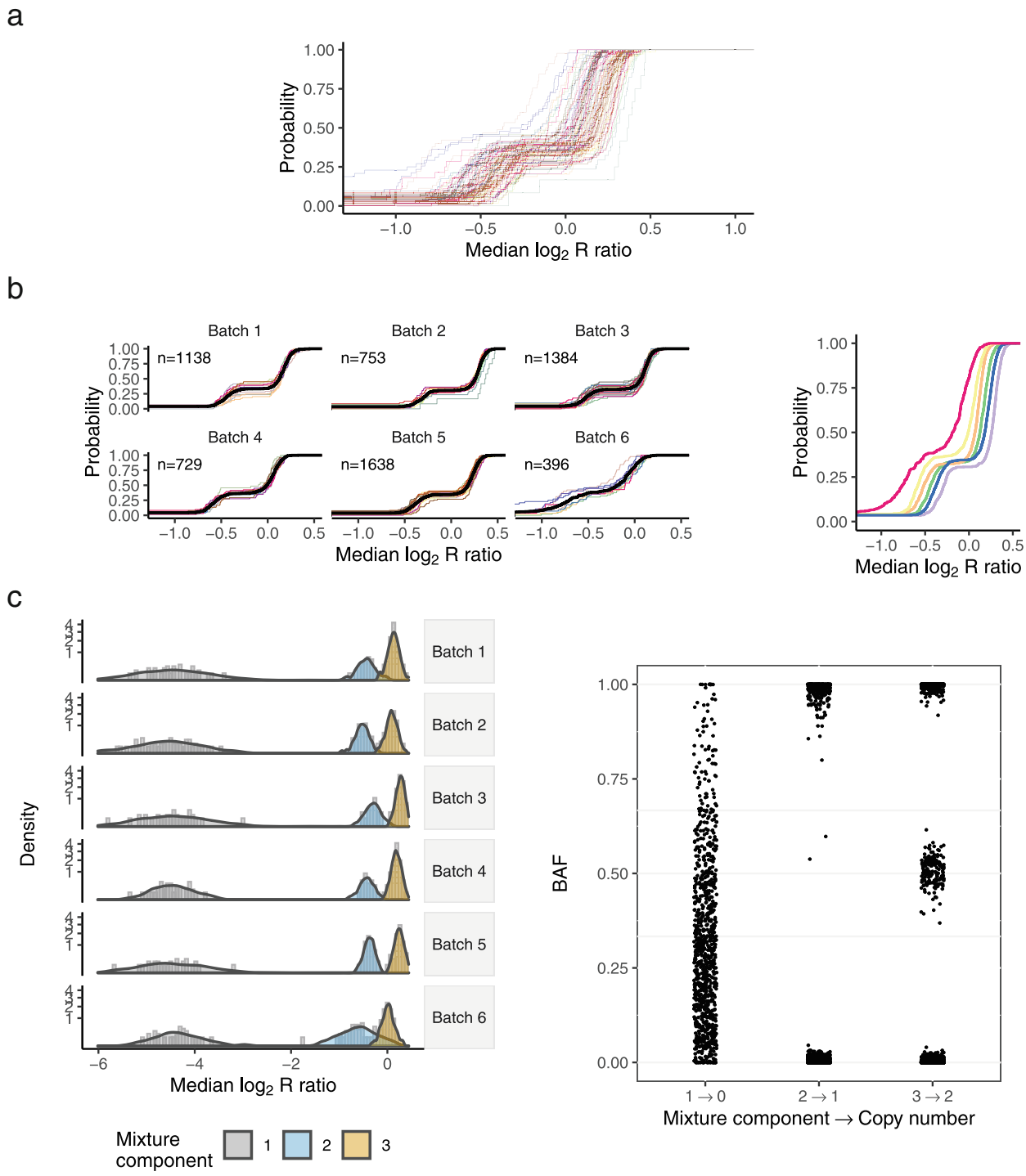


**Fig. 2** Identification of batch surrogates. **a** Plate-specific eCDFs of the average $\log_2$ *R* ratio for a region on chr5 (155,475,886-155,488,649bp). **b** The plate-specific eCDFs were grouped by Kolmogorov-Smirnov test statistics, forming batches. The batch-specific eCDFs after grouping plates (right). The eCDFs between batches typically differed by a location shift, though here Batch 6 also captured samples with higher variance. **c** Single- and multi-batch mixture models were evaluated at each CNP. Densities from the posterior predictive distributions overlay the histograms of the 3-component multi-batch model (left). Adjusted for batch, only three components were needed to fit the apparent deletion polymorphism. B allele frequencies were used to genotype the mixture components. The mapping from the mixture component indices to copy number is indicated by the arrows on the x-axis labels (right)

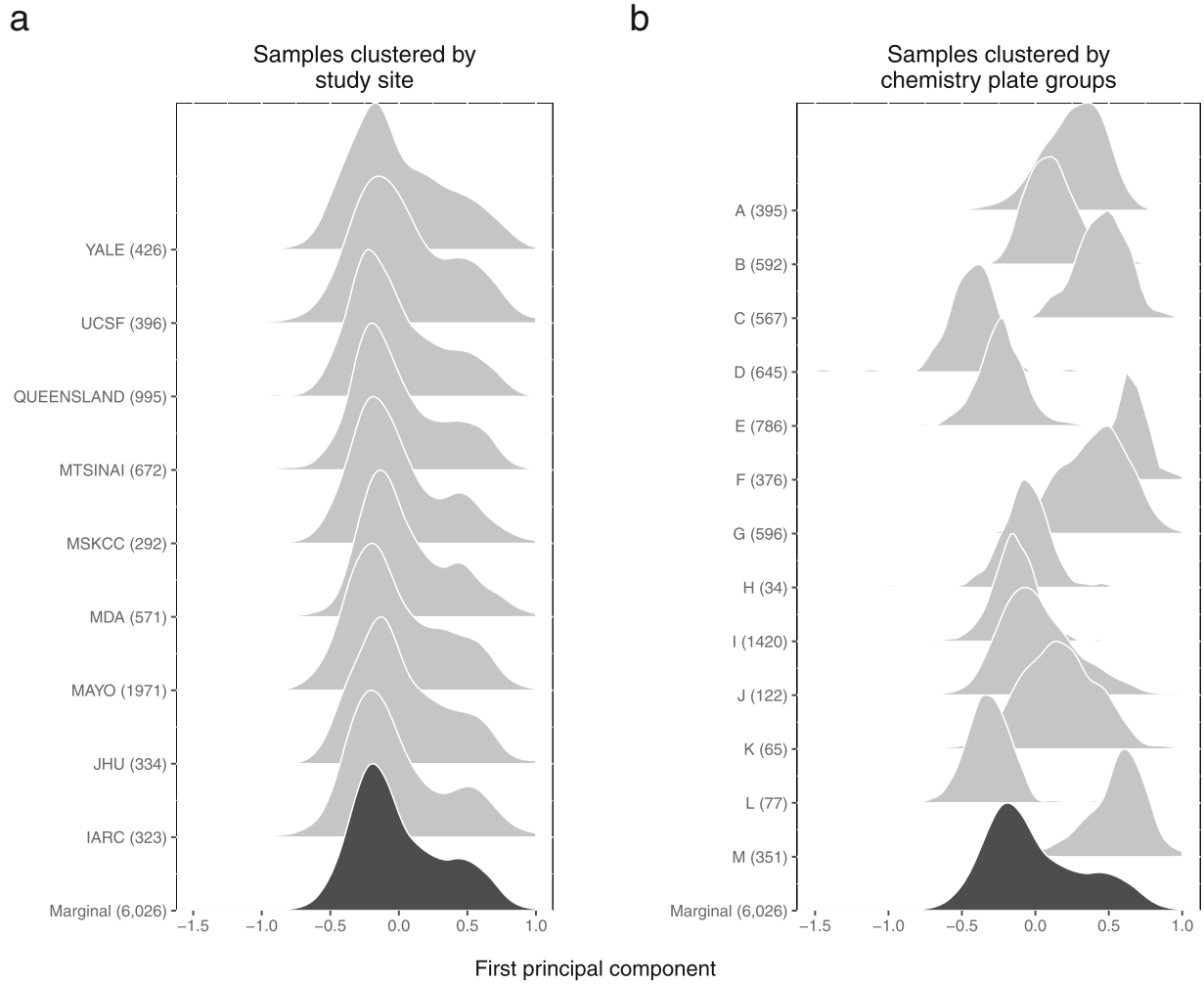


**Fig. 3** Study site does not capture the major sources of technical variation. Hybridization intensities were available for four probes in a CNP region on chr4 spanning 9,370,866 bp - 9,410,140 bp (CNP_051). Restricting our analysis to high quality samples, we used the first principal component (PC1) as a one-dimensional summary of the 4 x 6,026 matrix of $\log_2$ *R* ratios. The density of the PC1 summaries marginally (black) and stratified by study site (gray) are bimodal, suggesting a copy number polymorphism **a**. However, stratification of the PC1 summaries by grouping chemistry plates with similar eCDFs reveals an obvious batch effect (**b**). For example, chemistry plates in group E comprised of 786 samples originating from all nine study sites has a markedly different distribution than the 567 samples processed on group C chemistry plates

variances modeled hierarchically across batches (Fig. 2c). Following the probabilistic assignment of samples to mixture components, we genotyped the mixture components using the available BAFs at SNPs (Fig. 2c). From the 263 CNV regions, 25 regions contained samples with duplications, 132 regions contained samples with deletions, and 24 regions contained samples with deletions as well as samples with duplications. Allele frequencies from the genotyped duplications and deletions in the PanC4 controls were consistent with percentages reported in the 1000 Genomes Project. We identified a median of 17 additional CNVs per sample by the mixture model that were not identified by the HMM (Figures S4 and S5). On average, CNVs spanned 6 SNPs (interquartile range (IQR): 5-8) and were 12.6kb in size (IQR: 10.9kb - 17.6kb).

For 85 of the regions with deletions, small log ratios consistent with homozygous deletions appeared in a subset of the identified batches. Multi-batch models fit to these data require heavy-tails to accommodate the extreme observations and the resulting mixture components potentially capture different latent copy number states between the batches. Rather than exclude these observations, CNPBayes augments the observed data with simulated deletions. For the small number of individuals with likely germline homozygous deletions, their posterior probabilities can be interpreted as having been influenced by an empirically derived prior. Posterior probabilities for the remaining mixture components tend to be nearly equivalent to a model without augmentation fit to a dataset excluding the rare observations. For example,

the concordance of mixture component posterior probabilities comparing a model with augmentation to a model without augmentation that excluded 6 individuals with likely germline homozygous deletions at CNP_121 was near 1 (Figure S6).

#### *Comparison to other software*

Conceptually, our approach is most similar to CNVCALL [9] as Cardin *et al.* model one-dimensional summaries of CNV regions for each subject using a Bayesian hierarchical mixture of $t$ distributions [9, 31]. Below, we compare CNVCALL and CNPBayes at deletion and duplication polymorphisms in the PanC4 study that cover a range of data quality issues encountered in practice. When necessary, we performed a stratified analysis on the low and high quality samples. As CNVCALL does not interpret the copy number of the mixture components identified, we have labeled the copy number of their assigned components using the approach described for CNPBayes. For CNVCALL, we have used the first principal component as a one-dimensional summary for the CNV regions as recommended. Finally, we compare these methods to a set of simulations derived from HapMap samples where the true copy number is known.

*PanC4 study*

We performed a detailed analysis of four CNV regions in the PanC4 study (CNP_121, CNP_128, CNP_100, and CNP_240) that capture a range of data quality and copy number states (Figures S6–S9). For CNP_121, CNPBayes identified 5 batches in the high quality samples and a single batch in the low quality samples (Figure S6). A three component model was selected and the components were mapped to copy numbers 0, 1, and 2 from the BAFs as previously described, generating copy number frequencies of 9, 422, and 7167 (Hardy Weinberg equilibrium (HWE) $\chi^2_1$=1.15, p=0.28). Of the 7598 samples, 17 individuals were not called by CNVCALL, including 8 individuals with a missing $\log_2 R$ ratio in the CNV region and the 9 zero-copy individuals identified by CNPBayes (HWE $\chi^2_1$ = 6.18, p = 0.01). For the remaining 7581 samples, the posterior probabilities were highly concordant for both approaches. Similarly, CNP_128 is a deletion polymorphism. No batch effects were detectable in either the low or high quality data collections by CNPBayes (Figure S7). CNVCALL discarded 181 individuals at this locus, including 4 homozygous deletions identified by CNPBayes. The observed counts for copy numbers 0, 1, and 2 from CNPBayes were 4, 317, and 7277 (HWE $\chi^2_1$ = 0.08, p=0.77), while the corresponding counts for CNVCALL were 0, 303, and 7,114 ($\chi^2_1$ = 3.23, p=0.07). For CNP_100, CNPBayes identifies 6 batches in both the low- and hiqh-quality samples (Figure S8) and detects a duplication in the high quality samples but not for the lower quality data. CNVCALL did not identify any copy number alterations whether fit to all samples or when restricted to the set of high quality samples. BAFs among samples with the duplication identified by CNPBayes were highly consistent with three copies. Finally, for CNP_240 CNPBayes identifies copy numbers 0-3 at frequencies 2, 228, 5757, and 39 ($\chi^2_1 = 0.03, p = 0.87$ for copy numbers 0, 1, and 2) while CNVCALL identifies only hemizygous deletions ($n$=280) and diploid copy numbers ($n$ = 5647) ($\chi^2_1 = 3.47, p = 0.06$) (Figure S9). Overall, these analyses indicate that for regions where the signal to noise ratio is high, CNPBayes generates posterior probabilities of the latent copy numbers that are highly concordant with CNVCALL. Substantial differences in the two approaches arise for rare CNVs and for CNVs where the mixture components have greater overlap, often attributable to batch-to-batch differences in technical variation that are more flexibly modeled by CNPBayes. The CNPBayes assignment of relatively rare, large negative $\log_2 R$ ratios to a copy number zero state was consistent with expected frequencies of a deletion allele segregating in the population.

*Simulation*

To benchmark the sensitivity and specificity of this approach when the true genotypes were known, we extracted high quality data from a subset of HapMap phase III samples ($n$ = 990) processed on 16 chemistry plates and hybridized to Affymetrix 6.0 chips. A 109 kb region on chr 4 containing 1 SNP and 53 nonpolymorphic markers spans a deletion polymorphism with an allele frequency near 22%. We increased the level of difficulty for genotyping these samples by increasing the variance and/or shifting the location of the probe-level data in a subset of the chemistry plates. For each simulated dataset, we fit both CNVCALL and CNPBayes. While we did not provide the true batch labels to either method, CNPBayes estimated the batches from the plate surrogates. With no simulated batch effects, CNPBayes and CNVCALL had nearly identical performance with near perfect sensitivity and specificity (area under the receiver operator characteristic curve (AUC) $>$ 0.99). However, for simulated datasets with batch effects in the mean or variance, accuracy of CNVCALL decreased by an average of 25% while performance characteristics of CNPBayes remained qualitatively similar (Figure S10).

#### *Risk model for pancreatic cancer*

To evaluate whether changes in germline copy number effect pancreatic cancer risk, we fit a Bayesian logistic regression model at each CNV region. Uncertainty of the copy number assignment for each participant was incorporated in the regression model by sampling the integer copy number from a multinomial distribution parameterized by posterior probabilities from CNPBayes at each scan of the MCMC. As case-control status was unevenly distributed between the high and low data quality sample

collections ($\chi^2_1$=13.1, p=0.0003), the regression model included an interaction between copy number and data quality (Methods) as well as a single binary parameter $z_c$ multiplying both of these terms that allows the slopes to be exactly zero. The posterior mean of $z_c$ provides an estimate of the probability of an association with copy number (Fig. 4 and Table S2). Additional covariates included age, gender, and the first three principal components previously estimated from the SNP genotypes [20].

Genome-wide posterior probabilities of association between copy number and pancreatic cancer risk were near zero for most CNV regions (Fig. 4a). For five CNV regions with non-zero probabilities, we assessed the joint distribution of the regression coefficients for the high and low quality samples (Fig. 4b and c). Participants with two copies of the Tumor Supressor Candidate 3, *TUSC3*, had a 20% increased odds of pancreatic cancer compared to individuals with germline hemizygous deletions in this gene (90% credible interval (CI) for odds ratio: 1.01 - 1.39).

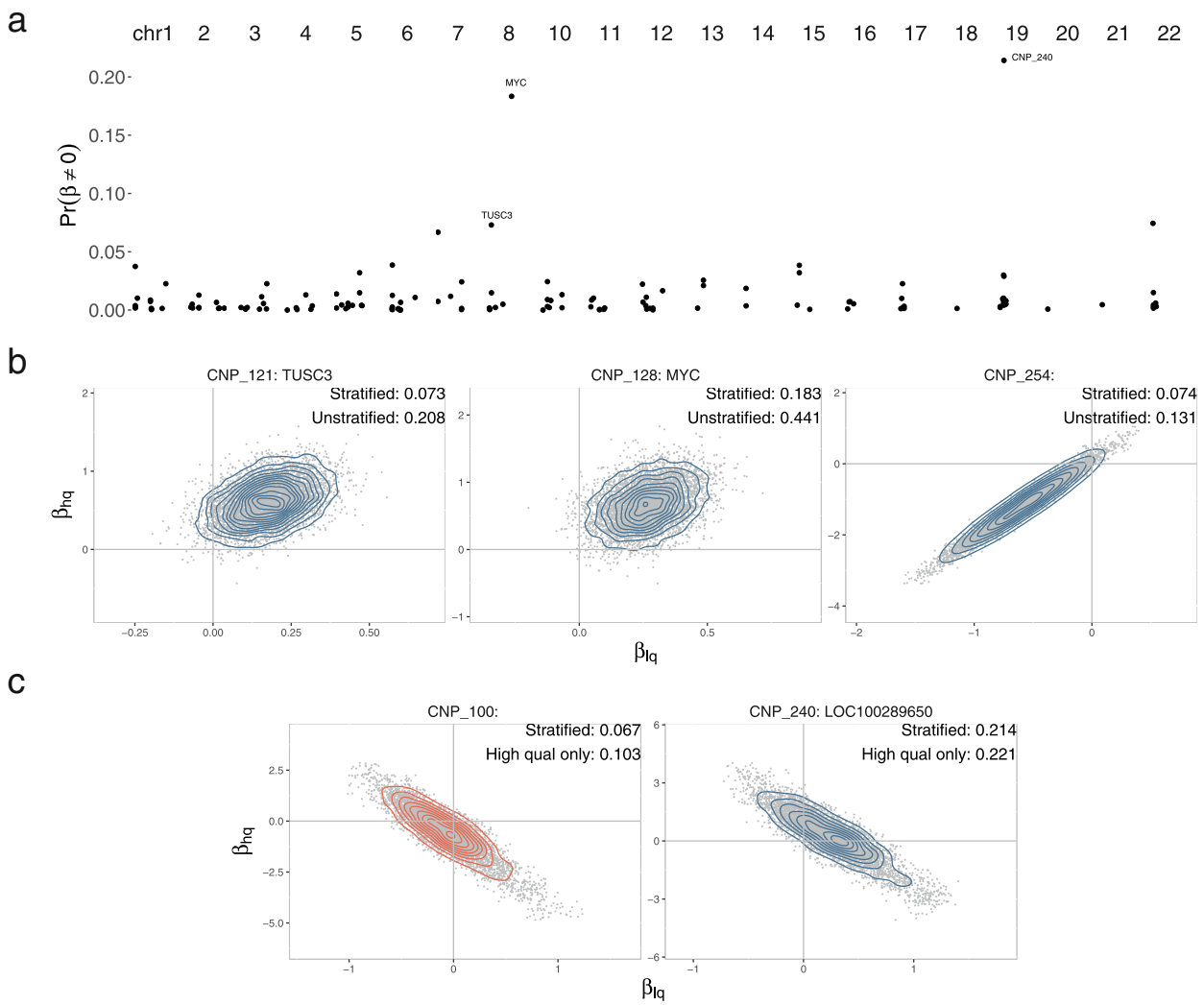


**Fig. 4** Bayesian regression models for pancreatic cancer risk. To incorporate uncertainty of the copy number assignment from the low-level data, the integer copy number was sampled from the subject-specific posterior probabilities provided by CNPBayes at each iteration of the MCMC. While batch effects on CNV inference were already accounted for in the low and high quality sample collections, an imbalance of the pancreatic cancer cases between these collections warranted a stratified model with an interaction between copy number and data quality and an indicator, $z_c$, multiplying these coefficients that allowed the slopes to be exactly zero. **a** Posterior probabilities of association from the stratified model for CNV regions across the genome. For regions where copy number inference was unaffected by data quality and associated with pancreatic cancer risk, regression coefficients for the low and high quality collections were positively correlated and the posterior mean of $z_c$ (upper right corner) increased in the more powerful unstratified analysis using all 7598 samples (**b**). By contrast, negatively correlated coefficients indicated an effect of data quality on CNV inference confirmed by visual inspection and the appropriate follow-up analysis and estimated probability of association was limited to the high quality sample collection (**c**)

While the direction of this effect is inconsistent with its putative role as a tumor suppressor, up-regulation of *TUSC3* and possible oncogenic roles for this gene have been reported in cancers including non-small cell lung cancer, colorectal, thyroid, and head and neck cancers [32–35]. Among non-coding regions, we found that deletions for a CNV region in 8q24 were associated with a reduced risk of pancreatic cancer (90% CI: 1.09-1.59). Chromosome 8q24 has been implicated in many cancers and is known to contain regulatory elements for the tumor oncogene *MYC* located at 128,748,315-128,753,680 bp [36]. We have previously demonstrated the association of SNPs in this region with an increased risk of pancreatic cancer [37, 38]. As copy number regression coefficients at CNV regions spanning *TUSC3* and near *MYC* were positive and highly correlated for both the low and high quality sample collections, an unstratified analysis using all 7598 participants doubled the posterior probability of association for these genes (Fig. 4b). Overall, our approach provides conservative measures of the association between copy number and pancreatic cancer risk across the genome, accounting for latent batch effects and copy number uncertainty separately for samples where data quality was more compromised.

## Discussion

We performed a genome-wide analysis of germline copy number variants in the largest study to date of pancreatic cancer, implementing approaches to correct for latent batch effects and risk models that incorporate uncertainty of the copy number estimates. As the batch effects we identified were likely related to differences in PCR efficiencies that can vary across the genome and between groups of samples processed on different chemistry plates within a single laboratory (not between study sites), we identified and adjusted for batch effects in a region-dependent manner in contrast to alternative methods. Using this approach, nearly 70% of CNV regions analyzed had multiple batches that were related to chemistry plates and not the individual laboratories that contributed samples.

Using the methods outlined in this study, we found that germline deletions of *TUSC3* and near *MYC* were more prevalent among participants without pancreatic cancer. Germline deletions of these genes have not been previously implicated in pancreatic cancer, though upregulation of expression of these genes have been implicated in some cancers in an apparent tissue-dependent manner. Although this study did not evaluate whether deletions at these loci were well tagged by neighboring SNPs, phasing the nearby SNPs would allow direct inference for whether variation in copy number is associated with pancreatic cancer risk among participants with the same SNP haplotype [39, 40]. While we evaluated copy number at both known and HMM-discoverable CNV regions for pancreatic cancer risk, more sensitive technologies for identifying smaller CNV regions with potentially rare germline CNVs among cancer patients are needed, and will not be well tagged by neighboring SNPs. Whether mosaic copy number alterations in hematopoietic cells could further modulate risk has not been evaluated [41–44].

Finally, we assumed an additive model for integer copy number and the log odds of cancer risk. Dominant and recessive mechanisms of genotype-phenotype associations are possible and the evidence for these models using Bayes factors could be averaged with weights reflecting our a priori beliefs.

## Conclusions

Statistical inference predicated on measures of abundance such as DNA copy number are highly susceptible to batch effects, and the sources of these effects are not generally known. As studies become increasingly large-scale with inevitable batch effects and heterogeneity in sample quality, the scalable approach provided by CNPBayes will be helpful for modeling unwanted technical variation and avoiding the potential confounding between batch effects and copy number when evaluating disease risk.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12885-020-07304-3.

**Additional file 1:** Figure S1: Median absolute deviation and autocorrelation of autosomal $\log_2 R$ ratios. Figure S2: Preprocessing and quality control analyses. Figure S3: Frequency of CNV regions with 1 to 7 batches identified by grouping the eCDFs of the $\log_2 R$ summaries. Figure S4: Number of additional CNVs identified from the Bayesian mixture model. Figure S5: Technical variation within and between samples obscures identification of hemizygous deletions. Figure S6: A deletion polymorphism at CNP_121. Figure S7: A deletion polymorphism at CNP_128. Figure S8: A duplication polymorphism at CNP_100. Figure S9: A CNV region with both deletions and duplications evident in the high quality samples. Figure S10: Performance of CNV detection methods on HapMap data.

**Additional file 2:** Supplemental Tables for Bayesian copy number detection and association in large-scale studies.

### Abbreviations

CNV: Copy number variant; HMM: Hidden Markov model; MCMC: Markov Chain Monte Carlo; PanC4: Pancreatic cancer case-control consortium; SNP: Single nucleotide polymorphism; ACF: Autocorrelation coefficient; eCDF: Empirical cummulative distribution function; K-S: Kolmogorov-Smirnov; MB: Multiple batches; SB: Single batch; MBP: Multiple batch pooled variance; SBP: Single batch pooled variance; HWE: Hardy Weinberg equilibrium; BAF: B allele frequency; PC: Principal component; IQR: Interquartile range; AUC: Area under the receiver operator characteristic curve

### Acknowledgments

We would like to thank Aravinda Chakravarti, Ann Oberg, Irene Orlow, and members of our laboratories for critical review of this manuscript.

### Authors' contributions

Conceptualization, APK and RBS; Methodology, SC, DM, JC, GR, IR, APK, and RBS; Contributing Data, PB, MD, SG, MGG, MMH, RJH, RCK, DL, LL, RN, SO, GP, KGR, HR; Formal Analysis, SC, DM, JC, JF, and RBS; Review of Manuscript, all

authors; Writing, SC, DM, JC, APK and RBS; Project Administration, APK and RBS; Funding Acquisition, APK. All authors have read and approved the manuscript.

**Funding**
This work was supported in part by the US National Institutes of Health grants 5R01CA154823, CA006973, and CA062924. The funders did not have any influence on any aspects of the study, including design, data collection, analyses, interpretation, or writing of the manuscript.

**Availability of data and materials**
The PanC4 study is available under dbGap accession number phs000206.v5.p3.

**Ethics approval and consent to participate**
Each participating study obtained informed written consent from participants and approval from their Institutional Review Board. This project was reviewed by the Johns Hopkins School of Medicine IRB.

**Consent for publication**
Not applicable.

**Competing interests**
S.C. and R.B.S. are founders of Delfi Diagnostics. R.B.S also holds equity in Delfi Diagnostics.

**Author details**
[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [2]Department of Oncology The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. [4]Genetics Section, International Agency for Research on Cancer, Lyon, France. [5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [6]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 10065 New York, NY, USA. [7]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, M5G 1x5 Toronto, Ontario, Canada. [8]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [9]Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA. [10]Department of Epidemiology, Cancer Prevention & Population Sciences, UT MD Anderson Cancer Center, 77030 Houston, TX, USA. [11]Department of Gastroenterology, Hepatology, and Nutrition Service, Memorial Sloan Kettering Cancer Center, 10065 New York, NY, USA. [12]Department of Gastrointestinal Medical Oncology, University of Texas MD Anderson Cancer Center, 77030 Houston, TX, USA. [13]Department of Chronic Disease Epidemiology, Yale School of Public Health, Yale Cancer Center, New Haven, CT, USA. [14]Population Health Department, QIMR Berghofer Medical Research Institute, 4029 Brisbane, Australia. [15]Department of Health Sciences Research, Mayo Clinic College of Medicine, 55905 Rochester, MN, USA.

Received: 21 February 2020 Accepted: 17 August 2020
Published online: 07 September 2020

**References**

1. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavaré S, Hurles ME. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. 2007;8(10):228. https://doi.org/10.1186/gb-2007-8-10-r228.
2. Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, Li HI, Qian H, Farinha P, Gascoyne RD, Marra MA. Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res. 2008;36(13):80. https://doi.org/10.1093/nar/gkn378.
3. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H., Bucan M, Maris JM, Wang K. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):126. https://doi.org/10.1093/nar/gkn556.
4. van de Wiel MA, Brosens R, Eilers PHC, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B. Smoothing waves in array CGH tumor profiles. Bioinformatics (Oxford, England). 2009;25: 1099–104. https://doi.org/10.1093/bioinformatics/btp132.
5. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):72. https://doi.org/10.1093/nar/gks001.
6. Leo A, Walker AM, Lebo MS, Hendrickson B, Scholl T, Akmaev VR. A GC-wave correction algorithm that improves the analytical performance of aCGH. J Mol Diagn JMD. 2012;14:550–9. https://doi.org/10.1016/j.jmoldx.2012.06.002.
7. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics. 2008;40(10):1253–60. https://doi.org/10.1038/ng.237.
8. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. Nature Genetics. 2008;40(10):1245–52. https://doi.org/10.1038/ng.206.
9. Cardin N, Holmes C, W.T.C.C.C, Donnelly P, Marchini J. Bayesian hierarchical mixture modeling to assign copy number from a targeted cnv array. Genet Epidemiol. 2011;35(6):536–548. https://doi.org/10.1002/gepi.20604.
10. Kumasaka N, Fujisawa H, Hosono N, Okada Y, Takahashi A, Nakamura Y, Kubo M, Kamatani N. Platinumcnv: a bayesian gaussian mixture model for genotyping copy number polymorphisms using snp array signal intensity data. Genet Epidemiol. 2011;35(8):831–44. https://doi.org/10.1002/gepi.20633.
11. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen M. J., Kirov G, Sullivan PF, Hultman CM, Sklar P, Purcell SM. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91(4):597–607. https://doi.org/10.1016/j.ajhg.2012.08.005.
12. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, Baras A, Overton JD, Habegger L, Reid JG. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2016;32(1):133–5. https://doi.org/10.1093/bioinformatics/btv547.
13. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):161. https://doi.org/10.1371/journal.pgen.0030161.
14. Costain G, Walker S, Argiropoulos B, Baribeau DA, Bassett AS, Boot E, Devriendt K, Kellam B, Marshall CR, Prasad A, Serrano MA, Stavropoulos DJ, Twede H, Vermeesch JR, Vorstman J. A. S., Scherer SW. Rare copy number variations affecting the synaptic gene DMXL2 in neurodevelopmental disorders. J Neurodevelopmental Disord. 2019;11:3. https://doi.org/10.1186/s11689-019-9263-3.
15. Kushima I, Aleksic B, Nakatochi M, Shimamura T, Okada T, Uno Y, Morikawa M, Ishizuka K, Shiino T, Kimura H, Arioka Y, Yoshimi A, Takasaki Y, Yu Y, Nakamura Y, Yamamoto M, Iidaka T, Iritani S, Inada T, Ogawa N, Shishido E, Torii Y, Kawano N, Omura Y, Yoshikawa T, Uchiyama T, Yamamoto T, Ikeda M, Hashimoto R, Yamamori H, Yasuda Y, Someya T, Watanabe Y, Egawa J, Nunokawa A, Itokawa M, Arai M, Miyashita M, Kobori A, Suzuki M, Takahashi T, Usami M, Kodaira M, Watanabe K, Sasaki T, Kuwabara H, Tochigi M, Nishimura F, Yamasue H, Eriguchi Y, Benner S, Kojima M, Yassin W, Munesue T, Yokoyama S, Kimura R, Funabiki Y, Kosaka H, Ishitobi M, Ohmori T, Numata S, Yoshikawa T, Toyota T, Yamakawa K, Suzuki T, Inoue Y, Nakaoka K, Goto Y-I, Inagaki M, Hashimoto N, Kusumi I, Son S, Murai T, Ikegame T, Okada N, Kasai K, Kunimoto S, Mori D, Iwata N, Ozaki N. Comparative analyses of copy-number variation in autism spectrum disorder and schizophrenia reveal etiological overlap and biological insights. Cell Rep. 2018;24:2838–56. https://doi.org/10.1016/j.celrep.2018.08.022.
16. Coe BP, Stessman HAF, Sulovari A, Geisheker MR, Bakken TE, Lake AM, Dougherty JD, Lein ES, Hormozdiari F, Bernier RA, Eichler EE. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat Genet. 2019;51:106–16. https://doi.org/10.1038/s41588-018-0288-4.
17. Huang K-L, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C., Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott A. D., Krassowski M, Cherniack AD, Houlahan KE, Jayasinghe R, Wang L-B, Zhou DC, Liu D, Cao S, Kim YW, Koire A, McMichael JF, Hucthagowder V, Kim T-B, Hahn A, Wang C, McLellan MD, Al-Mulla F, Johnson KJ, Network CGAR, Lichtarge O, Boutros PC, Raphael B, Lazar AJ, Zhang W, Wendl MC,

Govindan R, Jain S, Wheeler D, Kulkarni S, Dipersio JF, Reimand J, Meric-Bernstam F, Chen K, Shmulevich I, Plon SE, Chen F, Ding L. Pathogenic germline variants in 10,389 adult cancers. Cell. 2018;173: 355–37014. https://doi.org/10.1016/j.cell.2018.03.039.

18. Lucito R, Suresh S, Walter K, Pandey A, Lakshmi B, Krasnitz A, Sebat J, Wigler M, Klein AP, Brune K, Palmisano E, Maitra A, Goggins M, Hruban RH. Copy-number variants in patients with a strong family history of pancreatic cancer. Cancer Biol Ther. 2007;6:1592–9.
19. Willis JA, Mukherjee S, Orlow I, Viale A, Offit K, Kurtz RC, Olson SH, Klein RJ. Genome-wide analysis of the role of copy-number variation in pancreatic cancer risk. Front Genet. 2014;5:29. https://doi.org/10.3389/fgene.2014.00029.
20. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, Li D, Neale RE, Olson SH, Scelo G, Amundadottir LT, Bamlet WR, Bijlsma MF, Blackford A, Borges M, Brennan P, Brenner H, Bueno-de-Mesquita HB, Canzian F, Capurso G, Cavestro GM, Chaffee KG, Chanock SJ, Cleary SP, Cotterchio M, Foretova L, Fuchs C, Funel N, Gazouli M, Hassan M, Herman JM, Holcatova I, Holly EA, Hoover RN, Hung RJ, Janout V, Key TJ, Kupcinskas J, Kurtz RC, Landi S, Lu L., Malecka-Panas E, Mambrini A, Mohelnikova-Duchonova B, Neoptolemos JP, Oberg AL, Orlow I, Pasquali C, Pezzilli R, Rizzato C, Saldia A, Scarpa A, Stolzenberg-Solomon RZ, Strobel O, Tavano F, Vashist YK, Vodicka P, Wolpin BM, Yu H, Petersen GM, Risch HA, Klein AP. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat Genet. 2015;47:911–6. https://doi.org/10.1038/ng.3341.
21. Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. Ann Appl Stat. 2008;2(2):687–713.
22. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512–20.
23. Peel D, McLachlan GJ. Robust mixture modelling using the t distribution. Stat Comput. 2000;10(4):339–48.
24. Vounatsou P, Smith AFM. Simulation-based bayesian inferences for two-variance components linear models. J Stat Plan Infer. 1997;59(1): 139–61. https://doi.org/10.1016/S0378-3758(96)00093-6.
25. Lin TI, Lee JC, Ni HF. Bayesian analysis of mixture modelling using the multivariate t distribution. Stat Comput. 2004;14(2):119–30. https://doi.org/10.1023/B:STCO.0000021410.33077.10.
26. Chib S. Marginal likelihood from the Gibbs output. J Am Stat Assoc. 1995;90(432):1313–21. https://doi.org/10.1080/01621459.1995.10476635.
27. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7(4):457–472. https://doi.org/10.1214/ss/1177011136.
28. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C., Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X., He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui S. K. W., Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sall-e C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H., Kang L, Godbout M, Wallenburg JC, L'Archev-que P, Bellemare G., Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851–61. https://doi.org/10.1038/nature06258.
29. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the third international workshop on distributed statistical computing. 2003;124(125.10):1–10.
30. 1000 Genomes Project Consortium, Auton A, Brooks L, Durbin R, Garrison E, Kang H, Korbel J, Marchini J, McCarthy S, McVean G, Abecasis G. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
31. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N., Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews T. D, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JMC, Cooper JD, Dominiczak AF, Downes K, Drummond H. E., Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking L, Howard E, Howard P, Howson JMM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson T, Jolley JD, Jones IR, Jones LA, Kirov G, Langford CF, Lango-Allen H, Lathrop GM, Lee J, Lee KL, Lees C, Lewis K, Lindgren CM, Maisuria-Armer M, Maller J, Mansfield J, Martin P, Massey DCO, McArdle WL, McGuffin P, McLay KE, Mentzer A, Mimmack ML, Morgan AE, Morris AP, Mowat C, Myers S, Newman W, Nimmo ER, O'Donovan MC, Onipinla A, Onyiah I, Ovington NR, Owen MJ, Palin K, Parnell K, Pernet D, Perry JRB, Phillips A, Pinto D, Prescott NJ, Prokopenko I, Quail MA, Rafelt S, Rayner NW, Redon R, Reid DM, Renwick, Ring SM, Robertson N, Russell E, St Clair D, Sambrook JG, Sanderson JD, Schuilenburg H, Scott CE, Scott R, Seal S, Shaw-Hawkins S, Shields BM, Simmonds MJ, Smyth DJ, Somaskantharajah E, Spanova K., Steer S, Stephens J, Stevens HE, Stone MA, Su Z, Symmons DPM, Thompson JR, Thomson W, Travers ME, Turnbull C, Valsesia A, Walker M, Walker NM, Wallace C, Warren-Perry M, Watkins NA, Webster J, Weedon MN, Wilson AG, Woodburn M, Wordsworth BP, Young AH, Zeggini E, Carter NP, Frayling TM, Lee C, McVean G, Munroe PB, Palotie A, Sawcer SJ, Scherer SW, Strachan DP, Tyler-Smith C, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Gough SCL, Hall AS, Hattersley AT, Hill AVS, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Deloukas P, Duncanson A, Kwiatkowski D. P., McCarthy MI, Ouwehand W, Parkes M, Rahman N, Todd JA, Samani NJ, Donnelly P. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713–20. https://doi.org/10.1038/nature08979.
32. Gutiérrez VF, Marcos C. I., Llorente JL, Guervós MA, Iglesias FD, Tamargo LA, Hermsen M. Genetic profile of second primary tumors and recurrences in head and neck squamous cell carcinomas. Head Neck. 2012;34:830–9. https://doi.org/10.1002/hed.21824.
33. Gu Y, Wang Q, Guo K, Qin W, Liao W, Wang S, Ding Y, Lin J. Tusc3 promotes colorectal cancer progression and epithelial-mesenchymal transition (emt) through wnt/-catenin and mapk signalling. J Pathol. 2016;239:60–71. https://doi.org/10.1002/path.4697.

34. Gu Y, Pei X, Ren Y, Cai K, Guo K, Chen J, Qin W, Lin M, Wang Q, Tang N, Cheng Z, Ding Y, Lin J. Oncogenic function of tusc3 in non-small cell lung cancer is associated with hedgehog signalling pathway. Biochim Biophys Acta Mol Basis Dis. 2017;1863:1749–60. https://doi.org/10.1016/j.bbadis.2017.05.005.
35. Vašíčková K, Horak P, Vaňhara P. Tusc3: functional duality of a cancer gene. Cell Mol Life Sci CMLS. 2018;75:849–57. https://doi.org/10.1007/s00018-017-2660-4.
36. Grisanzio C, Freedman ML. Chromosome 8q24-associated cancers and MYC. Genes Cancer. 2010;1:555–9. https://doi.org/10.1177/1947601910381380.
37. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, Arslan AA, Beane-Freeman L, Bracci PM, Buring J, Canzian F, Duell EJ, Gallinger S, Giles GG, Goodman GE, Goodman PJ, Jacobs EJ, Kamineni A, Klein AP, Kolonel LN, Kulke MH, Li D, Malats N, Olson SH, Risch HA, Sesso H. D, Visvanathan K, White E, Zheng W, Abnet CC, Albanes D, Andreotti G, Austin MA, Barfield R, Basso D, Berndt SI, Boutron-Ruault M-C, Brotzman M, Bachler MW, Bueno-de-Mesquita HB, Bugert P, Burdette L, Campa D, Caporaso NE, Capurso G, Chung C, Cotterchio M, Costello E, Elena J, Funel N, Gaziano JM, Giese NA, Giovannucci EL, Goggins M, Gorman MJ, Gross M, Haiman CA, Hassan M, Helzlsouer KJ, Henderson BE, Holly EA, Hu N, Hunter DJ, Innocenti F, Jenab M, Kaaks R, Key TJ, Khaw K-T, Klein EA, Kogevinas M, Krogh V, Kupcinskas J, Kurtz RC, LaCroix A, Landi MT, Landi S, Le Marchand L, Mambrini A, Mannisto S, Milne RL, Nakamura Y, Oberg AL, Owzar K, Patel AV, Peeters PHM, Peters U, Pezzilli R, Piepoli A, Porta M, Real FX, Riboli E, Rothman N, Scarpa A, Shu X-O, Silverman DT, Soucek P, Sund M, Talar-Wojnarowska R, Taylor PR, Theodoropoulos GE, Thornquist M, Tjanneland A, Tobias GS, Trichopoulos D, Vodicka P., Wactawski-Wende J, Wentzensen N, Wu C, Yu H, Yu K, Zeleniuch-Jacquotte A, Hoover R, Hartge P, Fuchs C, Chanock SJ, Stolzenberg-Solomon RS, Amundadottir LT. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet. 2014;46:994–1000. https://doi.org/10.1038/ng.3052.
38. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D, Sun S, Yang F, Chen L, Wang J, Shah P, Cha SW, Aiyetan P, Woo S, Tian Y, Gritsenko MA, Clauss TR, Choi C, Monroe ME, Thomas S, Nie S, Wu C, Moore RJ, Yu K-H, Tabb DL, Fenya D, Bafna V, Wang Y, Rodriguez H., Boja ES, Hiltke T, Rivers RC, Sokoll L, Zhu H, Shih I-M, Cope L, Pandey A, Zhang B, Snyder MP, Levine DA, Smith RD, Chan DW, Rodland KD, Investigators C. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell. 2016;166:755–65. https://doi.org/10.1016/j.cell.2016.05.069.
39. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44. https://doi.org/10.1086/502802.
40. Scharpf RB, Mireles L, Yang Q, Köttgen A, Ruczinski I., Susztak K, Halper-Stromberg E, Tin A, Cristiano S, Chakravarti A., Boerwinkle E, Fox CS, Coresh J, Linda Kao WH. Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations. BMC Genet. 2014;15:81. https://doi.org/10.1186/1471-2156-15-81.
41. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q., Wang L-E, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, Blot WJ, Signorello LB, Ingles SA, Chanock SJ, Berndt SI, Le Marchand L, Henderson BE, Monroe KR, Heit JA, de Andrade M, Armasu S. M., Regnier C, Lowe WL, Hayes MG, Marazita ML, Feingold E, Murray JC, Melbye M, Feenstra B, Kang JH, Wiggs JL, Jarvik GP, McDavid AN, Seshan VE, Mirel DB, Crenshaw A, Sharopova N, Wise A, Shen J, Crosslin DR, Levine DM, Zheng X, Udren JI, Bennett S, Nelson SC, Gogarten SM, Conomos MP, Heagerty P, Manolio T, Pasquale LR, Haiman CA, Caporaso N, Weir BS. Detectable clonal mosaicism from birth to old age and its relationship to cancer. 2012;44(6):642–50. https://doi.org/10.1038/ng.2271.
42. Vattathil S, Scheet P. Haplotype-based profiling of subtle allelic imbalance with snp arrays. Genome Res. 2013;23:152–158. https://doi.org/10.1101/gr.141374.112.
43. Freed D, Stevens EL, Pevsner J. Somatic mosaicism in the human genome. Genes. 2014;5:1064–94. https://doi.org/10.3390/genes5041064.
44. Vattathil S, Scheet P. Extensive hidden genomic mosaicism revealed in normal tissue. Am J Hum Genet. 2016;98:571–578. https://doi.org/10.1016/j.ajhg.2016.02.003.


## Publisher's Note