Extended Experimental Procedures and Tables S1, S4, and S5

1 Helios Overview

Helios is a novel Bayesian algorithm that integrates information from heterogeneous data sources to identify driver genes within the complex landscape of somatic copy number alterations (SCNAs) of cancer.

Our approach proceeds in two steps (Figure 1A). First, our novel ISAR algorithm identifies regions of significant SCNA in the genome. Second, our novel Helios algorithm pinpoints the most likely driver gene within each such significant region.

ISAR estimates the significance of the alterations harbored by a marker based both on frequency and the background alteration rate at that genomic location. Instead of computing significance based on a genome-wide alteration rate (ref GISTIC), ISAR employs a local window to estimate the significance of the alteration with respect to its genomic location. Note that ISAR is completely decoupled from Helios and therefore could be substituted with any algorithm that identifies significant SCNA regions (Mermel, Schumacher et al. 2011), (Walter, Nobel et al. 2011), (Yuan, Yu et al. 2012, Yuan, Zhang et al. 2012) without any further modifications to Helios.

SCNA regions are typically large and it is therefore necessary to perform further analysis to pinpoint the target of the amplification. Unfortunately, in many cases, copy number alone is insufficient to pinpoint driver genes within altered regions. We utilize a Bayesian integrative model to incorporate cues from different genetic and genomic data types to distinguish driver from passenger genes. By combining features from different sources, signals that are individually insufficient, together could provide strong evidence of the oncogenic role of a gene.

The problem may be stated as binary classification in which a set of features is used to classify genes as either drivers or passengers. The standard setup for a classification problem requires a list of positive and negative examples, drivers and passengers in our system, to train the model. Unfortunately, the list of known oncogenic drivers is relatively small and strongly biased. Instead, Helios takes an unsupervised approach, harnessing a key property of drivers: drivers tend to have higher frequency of copy number aberration.

2 ISAR

We developed ISAR, an algorithm for detecting regions of statistically significant recurrent SCNAs in cancer. ISAR is based on the G-score metric, a significance measure of the aberration for each marker, which was originally defined in GISTIC (Beroukhim, Getz et al. 2007). Specifically, the G-score for a marker m is the summation of the copy number across samples that surpass an aberration threshold θ. Therefore, given the copy number for N samples, the G-score for a marker m in the case of amplifications is:

$$G^{AMP}(m) = \sum_{i=1}^{N} CN(m,i) \times I(CN(m,i) > \theta^{AMP}) \quad \textbf{Eq. 1}$$

Where CN(m,i) is the copy number of marker m in sample i and *I* is the indicator function.

State of the art algorithms like GISTIC2 compute a null distribution across the entire genome to estimate the significance of the alterations harbored by each marker. However, the alteration rate can strongly differ across genomic regions, due to features such as DNA secondary structure and DNA hypomethylation (De and Michor 2011). ISAR accounts for local differences in the alteration rate due to these and other unknown forces by scoring the significance of each alteration locally.

ISAR uses a local sliding window of constant size that moves along the chromosome, calculating the null distribution for each window. The use of a window allows the algorithm to estimate the local distribution of alterations and assign an accurate q-value to each marker based on its local distribution. Once the distribution has been computed in all windows within a chromosome, each genomic marker is associated with several overlapping windows. The algorithm takes a conservative approach by selecting the least significant q-value among the values computed for all overlapping windows containing the marker. By computing the significance locally, the algorithm is capable of identifying subtle events, such as a significant focal amplification within largely deleted regions, which would be missed if the background distribution for the whole genome were employed. For example, the pattern of alterations displayed by BCL2 in breast cancer becomes significant when you consider the background alteration levels of its region (Figure S1).

The algorithm is sensitive to the selection of the window size. Different window sizes are adequate for capturing events of different granularity: large windows tend to detect regions harboring large aberrations while small windows perform better in regions with small focal alterations. Therefore, ISAR is executed with several window sizes and the final score for each marker, denoted S-score, is the most significant q-value among the different window sizes.

$$S(m) = \max_{i \in W} -log10(qvalue_i(m)) \quad \textbf{Eq. 2}$$

Where W is the set of window sizes used and $qvalue_i(m)$, denotes the qvalue of marker m, using window size i.

Once the S-score has been computed for each marker, it is straightforward to define significantly altered regions. Each marker scoring above the user-defined peak threshold T_P is considered part of a region of alteration. These regions are extended to consecutive markers with a score above a user-defined region threshold T_R. Typical values for T_P are selected to match a q-value for the local window of 0.01-0.001 (T_P=2=-log10(0.01) - T_P=3=-log10(0.001)), while T_R is usually selected to match a slightly higher q-value, in the range of 0.1-0.01 (T_R=1=-log10(0.1) – T_R=2=-log10(0.01)).

2.1 Analyzing TCGA breast cancer data with ISAR

We used ISAR to analyze the amplification landscape of 785 breast cancer samples collected by the TCGA project using Affymetrix 6.0 SNP arrays (TCGA 2012). The segmented data was preprocessed to remove common CNVs and obvious technical artifacts. The total execution time was 4 hours on a laptop equipped with an Intel i7-2620M 2.7GHz processor and 8 GB of RAM. For this analysis, windows sizes of 4000, 5000, 6000 and 7000 markers were used and the window shift was set to 1/8 of the window size. T_P was set to 2 and T_R was set to 1. We used default values for the parameters shared with GISTIC (aberration thresholds=[0.15,-0.15], aberration caps=[2.0,-2.0], bin size=0.01).

A few post-processing steps are performed to ensure the quality of the detected regions. Adjacent regions closer than 250k base pairs were merged. Extremely small regions (size<10bp) were filtered out, as those are possibly due to artifacts. Regions that include the edges of the chromosomal arms were also discarded, as the statistical significance of the alteration in those regions is usually overestimated due to the difficulty of creating a null distribution for these locations of the genome

ISAR obtained 83 regions of significant SCNA (Table S1). On average these regions span 1256 Kb and contain 15 genes.

3 Helios: Bayesian integration for identifying drivers

Once the regions of significant SCNA have been detected, Helios uses an integrative Bayesian approach to rank the driver genes within each region. Helios uses a hierarchical Bayesian mixture model to distinguish drivers from passengers among the genes present in significantly altered regions. The unsupervised Bayesian algorithm discriminates driver genes (T=1) from passenger genes (T=0) by integrating the copy number alteration information (SCNA), with cues from different data sources (X). The hierarchical framework naturally separates these two components using the following model:

$$P(SCNA) = \sum_{t \in 0,1} P(SCNA|T = t)P(T = t|X) \qquad \textbf{Eq. 3}$$

Instead of predicting the classification as driver or passenger directly, the system learns by maximizing the likelihood of the observed copy number landscape (P(SCNA)). The graphical model has two components, one formalized as a classification $P(T = t|X)$ and a second that uses this classification to predict copy number. The assumption is that once the status of a gene as driver or passenger is given (T=0 or T=1), the frequency of alteration ($SCNA$) becomes independent from other predictive features (X). In summary, the approach separates the modeling of copy number ($P(SCNA|T = t)$) from other sources of information ($P(T = t|X)$) while focused on the predictive task of the observed copy number landscape ($P(CNA)$).

The algorithm iteratively fits a model for each part: $P(CNA|T = t)$ and $P(T = t|X)$ and updates the estimations (P(T)) for each gene taking both parts into account. The algorithm continues to

iterate until the model converges into a stable solution incorporating all the information into a single probability score for each gene.

Figure S2A shows the graphical model for Helios, where N genes are classified by combining the information from different data sources X and SCNA. w represents the parameters that control the integration of X, while λ parameterizes the influence of T on SCNA. In this model, when the values T_n for each gene n are given, the parameters for the different sources (W) and copy number (λ) are independent. This property makes it possible to fit the model efficiently using the Expectation Maximization (EM) algorithm.

3.1 Modeling copy number

A widely used approach to pinpoint driver genes within significantly altered copy number regions involves defining a minimal region of maximal alteration, called the peak region. While useful (Zender, Spector et al. 2006, Weir, Woo et al. 2007, Bass, Watanabe et al. 2009, Beroukhim, Mermel et al. 2010), this approach is insufficient for a number of reasons: (1) Even with a very stringent threshold, the minimal region of alteration can be fairly large and still contain multiple genes. (2) In many cases, the driver is not in the peak region, such as ADAM15 or BCL2 in the TCGA breast cancer dataset (3) While this approach assumes that each SCNA region is contains one driver gene, some regions may target several drivers while other regions might not contain a single driver gene because they target other regulatory elements (miRNA, lncRNA,…) or simply because they are the result of other forces that affect the rate of DNA breakpoints, such as genomic structure.

Helios considers the entire significantly altered region, but prioritizes the genes within each region using a Bayesian approach that makes explicit the uncertainty about the actual target/s of the SCNA. To achieve such prioritization, Helios uses additional sources of information to distinguish between genes which have equivalent copy number statistics and moreover the highest scoring gene need not be in the peak region. Finally, Helios can give a high score to more than one gene per region, or give low scores to all genes in a region.

Helios aims to model a distribution of SCNA that reflects the differences between driver and passenger genes, independent of the chromosomal region. However, in contrast to the subtle differences in SCNA within each altered region, the distribution of alterations differs dramatically between regions. Indeed, the median difference in G-score between genes in a region is significantly smaller (172) than the difference for genes across different regions (6405). Thus, without appropriate normalization, the G-score should not be used to prioritize drivers across regions. Ideally, instead of modeling the absolute number of alterations (which is dominated by the strongest alterations in the genome), we would like to model whether the gene is among the most altered genes in its own region (and therefore more likely to be the driver of that region). We therefore define a relative metric that measures the difference of each gene's G-score to the highest G-score in each region. That is, for a single gene g, we define the GSDist score as:

$$GSDist(g) = \max_{j \in region(g)} (Gscore(j) - Gscore(g)) \qquad \textbf{Eq. 4}$$

The most altered gene(s) in a region will have GSDist=0, while any other gene will have a positive value that indicates the "delta" in terms of G-score to the most frequently amplified gene in the region. Note that traditional approaches would use a threshold on this metric to make a hard decision on whether genes in the altered region are peak genes (Figure 2A). Instead Helios models this metric using two exponential distributions (one for drivers and one for passengers):

$$P(SCNA|\lambda_t) = \lambda_t e^{-\lambda_t GSDist} \quad \textbf{Eq. 5}$$

This model is based in the following intuition: the perturbation of driver genes provides a fitness advantage to cancer and therefore driver genes are likely to be among the most altered genes of their region, which translates into a GSDist distribution that exponentially decreases from zero with small variance. Passenger genes, on the other hand, have no evolutionary pressure to be selected for alteration and therefore can be modeled by a uniform distribution, which is approximated by an exponential distribution with large variance. This prior information on the variances of the two distributions is encoded into the algorithm through conjugate priors for λ_t. Considering a Gamma function for the prior distribution for λ_t, the posterior probability for λ_t belongs to the following Gamma distribution:

$$P(\lambda_t|SCNA) = Gamma(\alpha_t + \sum_g P_g(T=t)\,, \alpha_t * \beta_t + \sum_g P_g(T=t)GSDist(g)) \quad \textbf{Eq. 6}$$

3.2 Modeling additional sources of information

In most cases, the information extracted from copy number alone, is insufficient to pinpoint driver genes within altered regions. Helios overcomes this problem by incorporating cues from additional data sources that can facilitate the discrimination of driver genes from passenger genes. Helios's major strength is its ability to combine multiple weak pieces of evidence from heterogeneous data types to provide a strong indication of the oncogenic role of a gene.

The great challenge of data integration is to provide a unified framework to utilize all the data, despite the disparate nature of the features involved. In Helios, the information is unified by the function $P(T|X)$ that combines cues from all data sources into a single score. From the computational standpoint, $P(T|X)$ is a binary classifier that uses a set of features (X) to estimate whether the gene is driver (T=1) or passenger (T=0).

The features *X* are extracted from a diversity of data sources such as functional screens, gene expression and DNA sequencing. Features do not only help raise confidence in candidates, but also help to discard unlikely candidates. Indications like an unexpected frequency of point mutations or the oncogene addiction measured in shRNA screens can reinforce our confidence in driver candidates. While signals like the absence of variation in gene expression can help the algorithm discard passenger genes. We will describe the function $P(T|X)$ in greater detail after detailing some of the specific features *X.*

3.3 Features used in the Helios algorithm

The datasets are processed to extract features that can facilitate the identification of driver genes and distinguish these from passenger genes in SCNA regions. Some features, such as the significance of the number and location of point mutations harbored by a gene, are based on a single data type (DNA-sequencing), while others, like the score for oncogene addiction, are computed based on a combination of data types (gene expression and shRNA screens). In the following subsections we describe the features currently used in the Helios algorithms. Note that Helios provides a general framework where more features can be included as they become available and the algorithm can then automatically weigh them appropriately.

3.3.1 Sequence mutations

Driver genes can show a footprint of sequence mutations. This footprint consists of a higher recurrence of alterations, which in some cases may focus on specific locations such as certain functional domains or even a single base pair. In breast cancer and many other tumor types, the frequency of SCNAs for driver genes is significantly higher than the frequency of point mutations, with a handful of well-known exceptions such as PIK3CA and TP53.

We use MutSig (Banerji, Cibulskis et al. 2012) to compute the statistical significance of the recurrence of point mutations. MutSig tests the null hypothesis that the number of observed mutations in a gene can be attributed to a random background mutation processes, taking into account the bases covered as well as the length and composition of the gene. The computed p-value was log transformed to be used a feature for Helios.

3.3.2 Expression

Helios uses features extracted from RNA-Seq based gene expression to identify genes that are not expressed and those that, although expressed, do not seem to be driven by SCNA.

3.3.2.1 Expressed genes

Helios first uses RNA-Seq data to identify genes that are unlikely to be expressed in the tumors. Ramskold et al. (Ramsköld, Wang et al. 2009) concluded that RPKM measures can be employed to estimate whether a gene is expressed and estimated the percentage of genes that are expressed in different tissues. We compute the percentage of samples in which each gene is above this percentile and use it as a feature for Helios.

3.3.2.2 Association with alteration

The oncogenic activity of an amplified driver gene is expected to be reflected in the gene's mRNA dosage (Santarius, Shipley et al. 2010, Akavia, Litvin et al. 2011). We therefore anticipate the expression of the gene to be significantly higher in samples where the gene is amplified. We split the cohort into two groups, those samples in which the gene is amplified and those in which the gene is diploid and measure the association of amplification with expression using the Normal Approximation for the Wilcoxon rank sum test between these groups. As the driver mutation may only be operating in one subtype, this score is also computed for each subtype independently.

Genes that contribute to tumor development can also be overexpressed by different mechanisms in the absence of amplification. If those mechanisms prevail over copy number

amplification, the gene can present a lack of correlation between overexpression and amplification. We considered that those genes may show significant difference in expression between tumor types and therefore we measure the significance of the expression differences between subtypes, using the same test performed for association of amplification. Genes that do not show any significance in any of these two tests (p-value<0.05) are discarded and not scored.

3.3.3 shRNA

Although loss-of-function shRNA screens on tumor cell-lines are rapidly accumulating (Marcotte, Brown et al. 2012), (Cheung, Cowley et al. 2011), this strategy is still limited as an unbiased approach for the identification of tumor dependencies due to challenges such as off-target effects, low hairpin efficiency and the noise introduced by the stochastic nature of the pooled experiment (Kaelin 2012). Therefore, we calculate composite statistics combining the shRNA signatures with gene expression and breast cancer subtype information in order to elicit enhanced signal from this data before application of Helios.

3.3.3.1 Oncogene addiction score

The lack of reliable information about hairpin efficiency hinders the estimation of a ranking of gene lethality based on a single cell line. Therefore, to identify vulnerabilities, binary comparisons of hairpin dropout rates across several cell lines are usually performed. These comparisons are typically based on the classification of cell lines in tumor types or subtypes. Instead, we take a different approach based on the concept of oncogene addiction (Weinstein and Joe 2008): the perturbation of an oncogene produces dramatic changes in the cell, making it dependent on oncogene activity. Using gene expression as a proxy for the activity of an oncogene, we consider genes that show increased lethality when overexpressed to be likelier candidates. This idea has recently been used by Shao et Al. to discover the oncogene HNF1B (Shao, Tsherniak et al. 2013).

We score oncogene addiction by building a composite statistic reflecting the extent to which shRNA-depletion in a genome-wide screen is correlated with over-expression of the gene at baseline. The oncogene addiction score for a hairpin h with shRNA dropout S_h that targets a gene with expression profile $Exp_{target(h)}$ is the negative log likelihood of S_h given $Exp_{target(h)}$:

$$OA_h = -\log\left(P\left(S_h \mid Exp_{target(h)}\right)\right) = -\log(P(\epsilon)) \qquad \textbf{Eq. 7}$$

Where ϵ is the residual error vector of the shRNA dropout prediction vector $\widehat{S_h}$:

$$\epsilon = S_h - \widehat{S_h} = S_h - f(Exp_{target(h)}) \qquad \textbf{Eq. 8}$$

It is important to use a nonlinear regression for this prediction due to the strongly non-linear relationships observed in several bona fide cases, such as ERBB2 or FOXA1 (see Figure 2B-C). Therefore we model $f(Exp_{target(h)})$ using a linear ordering isotonic regression (Barlow, Bartholomew et al. 1972). We use the PAVA algorithm (Brunk 1955) to estimate the best fit for this regression.

We assume a Gaussian error $\epsilon \sim N(0, \sigma)$. Considering ϵ_i the error for cell line i, the oncogene addiction score for a hairpin is computed as:

$$OA_h = -\log\big(P(\epsilon)\big) = \prod_i -\log\left(\frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{1}{2}\epsilon_i\sigma\epsilon_i)\right) \quad \textbf{Eq. 9}$$

The variance for the error distribution σ is estimated from each hairpin independently. Note that many hairpins have extremely low variance, as there are a large number of inert shRNA hairpins and many others that target genes that are nonessential across all cell lines. To handle this situation, in instances where the hairpin's variance was smaller than the shRNA population variance, the latter was used as an estimate of the variance for the error distribution.

While the previous score is defined for a single hairpin, each gene is usually targeted by several hairpins. The final score for each gene is computed as the average of the best two scoring shRNA hairpins, as proposed by Marcotte et al. in the GARP score (Marcotte, Brown et al. 2012).

3.3.3.2 Subtype lethality score

For cancer types for which a molecular sub-classification is available, Helios includes a feature that scores the association between lethality and tumor subtype. We employ the same scoring scheme used for the oncogene addiction score, but in this case the predictor is a binary variable indicating the tumor subtype.

The oncogene addiction score and the subtype lethality score are not independent as in many cases (for example FOXA1, ESR1 and GATA3 in luminal breast cancer) overexpression and lethality are dominant across a whole subtype. We encode this dependency in the structure of the Bayesian network by introducing an additional intermediary node in the network that represents the overall lethality score for the gene and connects the two oncogene addiction and subtype lethality nodes to the final node (Supplementary Figure 2B).

3.4 Combining the features into a unified framework

The function $P(T|X)$ should be flexible enough to accommodate very diverse data types. At the same time, it should be constrained to avoid over-fitting. The key challenge is to combine these different features and weigh the relative contribution of each. Multivariate logistic regression is a common choice for classification problems with continuous features where over-fitting is a concern (Bishop 2006). However, the use of a simple logistic regression model is not well suited for this domain due to the strong non-linear nature of some of the features as well as the existence of dependencies between features.

Instead, we extended the logistic regression model by introducing additional layers of sigmoid functions (Figure S2B): each individual feature connects to a node representing a single sigmoid function and these are either combined into intermediary nodes and/or connected directly to a final node. The resulting classification is based on a final sigmoid function computed in this top

node. For example, as demonstrated in Figure S2B, the network contains two nodes related to shRNA which are joined into a single node that summarizes the lethality information for the gene. To avoid over-fitting, Helios uses Gaussian priors (Dan Foresee and Hagan 1997) for the parameters of this Bayesian network (**W**). These parameters serve multiple roles in determining how the signal is combined. At the simplest level, one can view the W as a way to weigh the different features, based on the importance of their contribution. Higher/lower values for **W** give their respective feature more/less weight in the final score. In addition, each feature goes through sigmoid nodes to adapt their input signal and the **W** parameters fit those as well, determining not only how much a feature contributes, but also thresholds defining when and how this contribution occurs.

3.5 Model learning

To classify genes as drivers (T=1) or passengers (T=0), the model (Figure S2A) requires fitting two sets of parameters: the parameters λ_t for the mixture of exponential functions that model the SCNA data and the parameters W of the Bayesian network that weigh and integrate other data sources. These sets of parameters are learned using the Expectation Maximization algorithm (EM) (Dempster, Laird et al. 1977), which iterates between two steps, optimizing the likelihood of the model in each step, until convergence. Note that when the classification T for the genes is given, the two sets of parameters become independent and they can therefore be learned independently. EM iterates between two stages that compute the estimates for T and the two sets of parameters respectively.

In the E-step the posterior odds for *T* are updated based on the prior odds $P(T = t|X, W)$ (computed by the Bayesian network that constitutes the integrative prior, Supplementary Figure 2B) and the likelihood ratio $P(CNA|T = t, \lambda_t)$ (computed by **Eq. 6**, which models the SCNA distributions):

$$\frac{P(T=1)}{1-P(T=1)} = \frac{P(CNA|\lambda_1)}{P(CNA|\lambda_0)} \times \frac{P(T=1|X,W)}{1-P(T=1|X,W)} \qquad \textbf{Eq. 10}$$

In the M-step the parameters λ_t and **W** are re-estimated using the updated values of P(T) computed in the previous E-step. Specifically:

- The parameters λ_t are recomputed as the expected value of their posterior Gamma distribution described in **Eq. 6**. Therefore the updated estimation is:

$$\lambda_t = \frac{\alpha_t * \beta_t + \sum_g P_g(T=t) GSDist(g)}{\alpha_t + \sum_g P_g(T=t)} \qquad \textbf{Eq. 11}$$

- The estimation of the optimal solution for the parameters **W** for the Bayesian network requires the computation of the Hessian matrix. To avoid computational overhead, we

use the Gauss-Newton approximation (Dan Foresee and Hagan 1997). We used the implementation provided by the function trainbr in Matlab's Neural Network Toolbox.

3.6 Initializing a starting point

The EM algorithm is not guaranteed to converge to the global optimum, but rather only to a local optimum. Therefore, a reasonably close initialization is key to achieve a good solution (McLachlan and Krishnan 2007). Helios needs to initialize the parameters for both the mixture of distributions that models the SCNA and the Bayesian network that models additional sources of information. Note that given the assignments of T, the two parts of the system are independent and all the parameters of each part can be learnt efficiently.

The initialization of the Bayesian network is based on the SCNA data. By selecting the most frequently altered gene in each region we can provide a good "first guess" for which genes are drivers. In practice, to initialize the parameters **W** for the Bayesian network that models $\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{X})$, we use a rough labeling of the subset of genes based on SCNA and point mutations. Genes that are significantly less aberrant than the top of their region (GSDist>150) are labeled passenger genes for the initialization process and the most altered genes in each region (GSDist=0) as well as those that are significantly mutated in sequence according to MusSig (p-value<0.01) are labeled driver genes. Using this binary assignment for P(T) Helios can learn the parameters for the network in the same way it would in an iteration of the M step. We therefore achieve an initial fitting of the parameters **W** and an initialization of $\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{X})$.

Helios then needs to initialize the parameters for the second part of the system, the mixture model that represents the SCNA information. As in the previous case, an assignment for P(T) would allow a fit of the parameters of the model ($\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$) using the same procedure performed in the M step of the EM algorithm. In this case, we use the current estimate of $\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{X})$ (obtained after the initial fitting of the Bayesian network) as the initial value of $\boldsymbol{P}(\boldsymbol{T})$**.**

4 Helios analysis of Breast cancer

4.1 Datasets used

We used the following public datasets:

- Primary tumor data from the TCGA Project (TCGA 2012): copy number Affymetrix 6.0 SNP arrays (n=785), Illumina HiSeq RNA sequencing (n=732) and whole-exome sequencing (n=507).
- Cell line shRNA screens (n=29) collected by Marcotte et al. (Marcotte, Brown et al. 2012).
- Cell line data from the Cancer Cell Line Encyclopedia (Barretina, Caponigro et al. 2012) for the cell lines screened with shRNA: copy number Affymetrix 6.0 SNP arrays (n=27) and messenger RNA Affymetrix U133 plus 2.0 arrays (n=27)

4.2 Copy number

Helios uses as candidate drivers the 1226 genes belonging to the 83 regions of significant SCNA identified by ISAR. The parameters for the Gamma priors for the copy number model are

set to $\beta_1 = 125$ and $\alpha_1 = 1000$ for driver genes and $\beta_0 = 300$ and $\alpha_0 = 4000$ for passenger genes, which emphasizes a smaller GSDist for drivers. Different values were also tested for these parameters without displaying any significant impact in the results of the analysis.

4.3 Subtype Classification

Breast cancer is a heterogeneous disease with different molecular subtypes. Despite the discrepancies between different molecular classifications that have been proposed in the literature (Perou, Sorlie et al. 2000),(Prat and Perou 2011), (Curtis, Shah et al. 2012), all authors agree on the existence of two main molecular subtypes of breast cancer: luminal and basal. Therefore we considered these two subtypes in our analysis. The subtypes for the cell lines were obtained from Marcotte et al. (Marcotte, Brown et al. 2012). The primary tumors are classified into subtypes using receptor status recorded in the clinical annotations, where estrogen and progesterone negative tumors are considered basal and any other tumor is considered luminal.

The association of alteration with each subtype was estimated using the G-score. The G-score for the samples in each subtype was calculated and the significance of this score was estimated by permutation testing: 10000 random permutations of the data are generated and the G-score for the subtype in each permutation was compared against the subtype G-score. Genes that displayed a significant p-value for this score (<0.01) were deemed subtype-specific.

Features for Helios were computed for the whole cohort and for each subtype independently. For alterations that are subtype-specific, the value of the feature computed for the subtype associated with the alteration was employed. For alterations that do not display association with subtype, the most significant value between the three computed (all cohort, basal and luminal) was used.

4.4 Features

4.4.1 Sequence mutations

We use MutSig (Banerji, Cibulskis et al. 2012) to compute the statistical significance of the recurrence of point mutations in the 507 samples sequenced by TCGA. The MutSig analysis was obtained from the GDAC TCGA pipeline. Specifically, version 2011112800.0.0 (MutSig v1.5) was employed.

4.4.2 Expression

Helios uses several features extracted from RNA-Seq based gene expression from 732 patients collected by the TCGA. The RPKM processed data was obtained from the GDAC pipeline.

We use the procedure described in Section 3.3.2 to compute the feature that estimates whether a gene is expressed. The distribution for this feature displayed strongly bimodality and based on this bimodality we decided not only to use it as a feature for Helios in this case, but instead to filter out genes based on this criteria. Genes that had RPKM values above the threshold in less than 30% of the samples were filtered out and removed from consideration.

4.4.3 shRNA

For our analysis we use the Breast cancer cell lines from the shRNA screen collected by Marcotte et al. (Marcotte, Brown et al. 2012). We use the computed shARP score as defined by the authors as a measure of lethality for each cell line. The shARP score was median normalized and then standardized using the deviation of the positive values. Because RNA-seq data was not available, considering the inherent noise of mRNA microarrays and that part of the genome was not measured by the array, we also computed the score using copy number instead of mRNA and the most significant score out of the two was selected. For genes not assayed in the shRNA screen, we used the mean value of the features across all genes to fill in the missing values.

5 Convergence

We executed Helios on the dataset following the initialization described in Section 3.5. The algorithm converged to a stable solution after 25 iterations as shown in Figure S3A-B.

6 Stability of Helios

We performed 100 runs randomly and uniformly sub-sampling 95% of the samples in each execution. The percentage of runs in which a region is called by ISAR depends clearly on the ISAR score (Figure S3C). Peaks with S-score above 3 were called on average in 98.30% of the executions. We observed that although displaying more private focal mutations, several regions with scores below 3 contained known driver genes such as FOXA1, PIK3CA, GAB1, MYB or NOTCH3. We decided to lower the threshold to 2 considering that the integration of other types of data performed by Helios would increase the confidence in the presence or absence of driver genes in these regions.

We tested Helios' stability by comparing gene rakings across sub-sampled runs using Pearson correlation. Figure S3D shows the histogram of pairwise correlations across runs. Helios demonstrated exceptional robustness, displaying an average correlation between Helios scores across executions of 0.96 and with the lowest correlation being 0.81.

7 Performance of data integration

To assess performance, a gold standard set of 330 drivers was compiled from the following sources:

- The set of known amplified oncogenes from Beroukhim et al. (Beroukhim, Mermel et al. 2010)
- The set of genes related to Breast cancer according to the University of Copenhagen DISEASES database (Frankild and Jensen) with score greater than 2.5. This list includes both oncogenes and tumor suppressors. We filtered out genes categorized as tumor suppressors according to Uniprot (Consortium 2013).

We compared Helios to alternative methods both in terms of the number of gold standard genes captured and the p-value of the hypergeometric enrichment of gold standard genes in the predicted set.

We first compared the performance of Helios against the simple criteria of selecting the gene with the largest G-score per region. Out of the 118 genes that were top of their region (some regions have more than a single gene at the top), 15 are annotated in the gold standard. Helios identifies the same number of annotated genes but achieved this with only 64 genes that had score greater than 0.5 selected. Thus, it achieves the same sensitivity but significantly increases the specificity, increasing the enrichment p-value from 8.40E-10 to 8.16E-14.

We then compared the performance of the integrative approach against the simple selection of candidates within SCNA regions based on each data source individually. We also compared selecting genes solely based on SCNA, choosing those that were clearly more altered than any other gene in the region. We selected top altered genes in their region where the second altered gene had a GSDist>150. Similar results were obtained with thresholds between 100 and 200. Figure 2F shows the performance of Helios compared to both the analysis of each independent data source and the union of genes captured by all independent data sources. Helios clearly outperformed each individual data source and the union of all data sources both in terms of sensitivity and specificity. This result exemplifies the power and benefits of data integration, in which all the information is considered simultaneously in a unified model that leverages subtle signals to achieve better performance than what would be obtained by analyzing each data source independently.

To assess the contribution of each data source to the performance of Helios, the algorithm was executed excluding each of the data sources (shRNA,expression and sequence mutations) and the enrichment of the genes in the gold standard was assessed with GSEA (Subramanian, Tamayo et al. 2005). Figure S3E shows that the contribution of functional screens is greater than the one of any other feature, but all features contribute to the performance.

8 Evaluation of Helios's accuracy

Here we summarize the different ways in which we evaluated Helios's performance. Focusing on the Helios score itself, we found that 9/10 (90%) of the highest Helios scoring genes are well known breast cancer oncogenes.

However, Helios's was designed to rank genes within a region and therefore the better way to evaluate Helios is to ask "*How often does Helios correctly rank the driver at the top of its region*". The problem is that we don't know what the correct answer is for the majority of the regions and therefore used the ISAR score to select the 17 most significantly amplified regions for evaluation. Selecting these 17 regions has two advantages: (1) Any identified drivers will be involved aberrant in the largest number of patients. (2) The ISAR score is independent of the Helios score and this gives us a wide range of Helios scores at the top of their respective regions (0.36 to 0.95).

Of the top 17 regions, we could only test 14. One region had no high scoring gene and indeed this region contained the oncomir mir21. In addition, we failed to clone the top gene for 2 otherregions, both lacking any known oncogene, high scoring or other. Among the remaining 14 regions, 6 contained known oncogenes in breast cancer (ERBB2, CCND1, ZNF217, MYC, FGFR2 and IGF1R) and were therefore not considered further. The remaining 8 regions contained no known oncogenes in the entire region and were thus subject to further experiments. In some cases we experimentally tested more than one gene in a region, when the region contained more than one high scoring gene, resulting in 12 genes selected for validation. Considering all of the experimentally tested genes, 10/12 (83%) successfully validated. Considering only the top scoring gene in each region, 7/8 (88%) successfully validated.

9 Performance comparison with other methods

We compared the capabilities of Helios against the state of the art algorithms GISTIC2 (Mermel, Schumacher et al. 2011) , Gaia (Morganella, Pagnotta et al. 2011) and DiNAMIC (Walter, Nobel et al. 2011) (Figure 3A). GISTIC2 identified 30 peaks containing 452 genes (TCGA 2012), out of which 17 are annotated in the gold standard, yielding a hypergeometric enrichment P-value of 1.2E-3. 16 of the 83 top Helios genes for each of the 83 amplified regions discovered by ISAR are annotated, indicating an enrichment p-value of 4.71E-12. The capability of the Helios score to discriminate drivers becomes even more evident if we further select only the 64 genes with a Helios score greater than 0.5, yielding a hypergeometric enrichment of 8.16e-14. The other two recently published methods tested, Gaia and DiNAMIC, achieved poor specificity compared to Helios, as reflected by their enrichment p-value: 7.7E-2 and 9.9E-2 respectively.

10 Module analysis

We used a modified version of Multi-Reg (Danussi, Akavia et al, Cancer Research 2013) to identify potential targets of RSF1. We ran Multi-Reg once on the Basal samples and once on the Luminal samples. For each sample type, we unified all modules generated by Multi-Reg that were associated with RSF1 into two modules - genes induced and genes repressed by RSF1 (see Figure 6A and Figure S6). All other parameters of Multi-Reg were as described in the original article.

We downloaded the C2 subcomponent of the MSigDB signature database version 3.1 from http://www.broadinstitute.org/gsea/msigdb/index.jsp on May 9th, 2013 (Subramanian, Tamayo, et al. 2005, PNAS 102, 15545-15550). We ran hypergeometric enrichment using the Genatomy software downloaded from http://www.c2b2.columbia.edu/danapeerlab/html/genatomy.html (see Figure 6A, Figures S6).

11 Experimental methods

11.1 Cell culture and reagents

Cell lines were obtained from the American Type Culture Collection (Manassas, VA, USA). Both the human normal breast epithelial MCF10A cell line and the MCF10A-TM (Pires et al., 2012) cell line were grown in DMEM/Ham's F-12 media (Corning, 10-092-CV) supplemented with 5% horse serum (Invitrogen, 16050-122), 20ng ml^{-1} EGF (Sigma, E9644), 1ng ml^{-1} Cholera Toxin (Sigma, C8052), 10mg ml^{-1} Insulin (Sigma, I9278), 100mg ml^{-1} Hydrocortisone (Sigma, H0396) and 1% of Penicillin-Streptomycin (Life Technologies, 15140-122). The mouse mammary epithelial Comma-1D cell line was grown in DMEM/Ham's F-12 media (Corning, 10-092-CV) supplemented with 5% fetal serum (Invitrogen, 16050-122), 10ng ml^{-1} EGF (Sigma, E9644), 5mg ml^{-1} insulin (Sigma, I9278), 200mg ml^{-1} Hydrocortisone (Sigma, H0396) and 1% of Penicillin-Streptomycin (Life Technologies, 15140-122). The MDA-MB 415 cell line was cultured in DMEM media (Sigma, D6429) supplemented with 15% fetal serum, bovine insuline 100mg ml^{-1} (Sigma I0516) and 100mg ml^{-1} of 85% glutathione (Sigma G-6013). The MDA-MB 361 and MDA-MB 453 cell lines were cultured in L-15 media (ATCC 30-2008) supplemented with 1% of Penicillin-Streptomycin (Life Technologies, 15140-122) and 20% and 10% of fetal serum respectively.

For generation of cell lines overexpressing RSF1 or other genes, cells were plated at 60% confluence in a 6 well plate and after 24 hours, were infected with lentivirus expressing the different construct plasmids. Media containing lentivirus was replaced in 12h for fresh media. After that cells were re-infected for other 12h. Cells were grown in fresh media for 24h and selected with the appropriate drug. Alternatively, to generate MDA-MB 453 deficient in RSF1, cells where infected with lentivirus espressing doxycycline-inducible pTRIPz shRNA against RSF1 (from Open Biosystems V3THS_341214, V3THS_341216, V3THS_341217) and selected with the puromycin (2ug/mL).

The colony formation assay in semisolid media was performed in 6 well plates. First, a layer of 2 mls of 0.6% agar (Fisher #9002-18-0) in regular MCF-10A media was placed at the bottom of each well and allowed to undergo gelification. Then, a layer of 2 mls of 0.3% agar containing 5,000 cells was seeded on top of the bottom agar layer and allowed to form a gel. Finally, 1 ml of regular MCF-10A media was placed covering the agar. The colonies were allowed to form for 1 month. After this period 2 mls of MTT solution (Sigma #M5655) at 0.5mgr/ml was used to stain the colonies. A minimum of 6 replicas per gene were plated. The number of colonies was independently evaluated by two researchers.

11.2 DNA constructs and gene cloning strategy

cDNAs from genes of interest were obtained from hORFeome V8.1 and CCSB Broad Libraries. The cDNAs from genes not contained in such libraries were amplified from regular RNA extracted from MCF-10A and retrotranscribed to cDNA with oligo-dT by PCR using specific primers bearing restriction sites (BglI, BamHI, XhoI or EcoRI). Briefly, RNA was extracted using RNeasy extraction Kit following the manufacturer‘s recommendations (QIAGEN #74106). 0.5 micrograms of total RNA was converted into cDNA using the High Capacity cDNA Reverse Transcription Kit (Roche #4368814) according to manufacturer's instructions.
Amplified products were cloned into a modified *pLPCX* vector (mod-pLPCX) in which we inserted an IRES2-EGFP from the donor pIRES2-EGFP.

Once the cDNAs were cloned into the recipient mod-pLPCX vector, all constructs were sequence verified by Sanger sequencing.

Primers used are specified in Supplementary Table 4.

The Firefly Luciferase expressing construct used to generate MCF10A-TM cell line was a gift from Jan Kitahewski (Columbia University, NY, NY).

11.3 Determination of RNA levels and RT-PCR analysis

Total RNA was extracted from cells according to manufacturer's instruction using RNeasy Kit (Qiagen, 74106). 2 µg of RNA was used as a template for reverse transcription using random primers. Reverse Transcription of RNA was performed as directed using TaqMan MicroRNA Reverse Trancription Kit (Applied Biosystems, 4366596) in a 20µl volume. 2ml RNA was used for template for RTQ-PCR using FastStart SYBR Green Master (Roche, 04 673 492 001). Reactions were performed in triplicate. The program reaction was: AmpliTaq activation 95ºC for 3 minutes, denaturation 95ºC for 10 seconds, and annealing/extension 60ºC for 30 seconds (repeat 40 times). Triplicate Ct values were further analyzed ($2^{-\Delta\Delta CT}$) by normalizing to an endogenous reference gene (β-actin). Results are presented as the relative mRNA amount compared to the samples transduced with control empty vectors. Forward and reverse primer sequences are specified in Supplementary Table 5.

11.4 Protein extracts and Western Blot

Cells were washed with cold PBS and lysed with EZ lysis buffer (1M Tris pH7, 50% glycerol, 20% SDS, 1mM ortovanadate, 1mM sodium fluoride and 1mM phenylmethylsulfonyl fluoride). Protein concentrations were determined by the Protein Assay Kit (Bio-Rad #500-0006). Equal amounts of proteins were subjected to SDS-PAGE and transferred to nitrocellulose membranes (GE Healthcare #10401197). Non-specific binding was blocked by incubation with TBST (20 mM Tris-Hcl pH7.4, 150 mM NaCl, 0.1% Tween-20) plus 5% of non-fat milk. Membranes were incubated with the primary antibodies overnight at 4ºC and for 1 hour with secondary HRP-conjugated antibodies at room temperature (Amersham #NA9350V, #NA931V and #NA934V). Signal was detected with Lumi-Light Western Blotting Substrate (Roche #12015200001 and #12015196001).

The antibodies used in this study were the following: human RSF1 (GenTex #GTX62703), CCND1 (BD Pharmingen #556470) and b-Actin (USBiological #A0760-40),)

11.5 Tumorigenicity in mice

Animal maintenance and experiments were performed in accordance with the animal care guidelines and protocols approved by Columbia University animal care unit. For the Comma-1D cell line, 21 days old female NOD.CB17-Prkdcs SCID mice (Harlan) mice were injected with $5x10^5$ cells, resuspended in PBS, into a fat mammary gland. For the MDA-453 cell line, eight-weeks old female NOD.CB17-Prkdcs SCID mice (Harlan) mice where injected with $5x10^6$ cells, resuspended in 1:2 Matrigel (BD Biosciences) plus normal growth media, into a fat pad mammary gland. Doxicyclin was added to drinking water at a final concentration of 2.0 mg/mL. Tumor growth was monitored twice a week with callipers at the site of injection. Animals were sacrificed as soon as tumor size reached 1.5 cm diameter.
In the experimental metastasis assays, eight-weeks old female NOD.CB17-Prkdcs SCID mice (Harlan) where injected with $5x10^6$ cells, resuspended in PBS, via the tail vein. To measure the luciferase intensity of injected cells, 2.25 μg of luciferin was injected intravenously through the tail and luciferase activity was assessed 5 minutes after luciferin injection using a IVIS Spectrum pre-clinical in vivo Imaging System (PerkinElmer, IVISSPE) machine. The presence of established metastases was visually confirmed after euthanizing the mice.

11.6 H&E and Immunohistochemistry

Mammary glands were fixed in formalin (Fischer #175) for immunohistochemistry (IHC) analysis. Formalin-fixed paraffin-embedded samples were firstly heated at 100ºC for 3 minutes on a heat block to melt the paraffin. Subsequently, samples were deparaffinised by a serial incubation with xylene for 3 minutes, 100% EtOH for 3 minutes, 95% EtOH for 3 minutes and distilled water for 2 minutes. Peroxidase inactivation and antigen retrieval were achieved by incubating samples in 1% H2O2 for 15 minutes at room temperature and incubating slides with citric buffer (2mM citric acid, 8mM sodium citrate) in a steamer for 30 minutes. Samples were washed twice in PBS for 5 minutes and incubated in 10% whole goat blocking serum diluted in 2% BSA-PBS for 30 minutes at room temperature. Thereafter, samples were incubated in primary antibody [1:200 Ki67 (Abcam #ab15580), 1:300 Anti-Cytokeratin 18 antibody (Abcam #ab668), 1:1000 Anti-Cytokeratin 5 (Covance #PRB-160P) and 1:500 Cleaved Caspase-3 (Cell Signaling #9664)] and diluted in 2% BSA-PBS+0.01% sodium azide for 2 hours at room temperature.

Samples were then washed in PBS and incubated in 1:500 biotinylated anti-Rabbit IgG made in goat diluted in 2% BSA-PBS for 30 minutes at room temperature. Afterwards, samples were washed and exposed to peroxidase substrate (Vector Laboratories #PK-6100) for 30 minutes at room temperature and subsequently permeabilized with PBS-0.5% Triton. Thereafter, samples were incubated in chromogen 3,3' Diaminobenzidine (DAB) and then washed in distilled water and counterstained. Counterstaining was performed by treating samples with hematoxylyn for 1 second, dipped in 1% Hydrochloric acid and finally washed in ammonia water for 1 second. Finally, dehydration was performed by incubating samples in 95% EtOH for 2 minutes, 100% EtOH for 2 minutes and xylene for 4-5 minutes and ultimately mounted with a coverslip.

Supplemental References

Akavia, U. D., et al. (2011). "An Integrated Approach to Uncover Drivers of Cancer." Cell **143**(6): 1005-1017.

Banerji, S., et al. (2012). "Sequence analysis of mutations and translocations across breast cancer subtypes." Nature **486**(7403): 405-409.

Barlow, R. E., et al. (1972). Statistical inference under order restrictions. New York, Wiley.

Barretina, J., et al. (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." Nature **483**(7391): 603-307.

Bass, A. J., et al. (2009). "SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas." Nat Genet **41**(11): 1238-1242.

Beroukhim, R., et al. (2007). "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma." Proceedings of the National Academy of Sciences **104**(50): 20007-20012.

Beroukhim, R., et al. (2010). "The landscape of somatic copy-number alteration across human cancers." Nature **463**(7283): 899-905.

Beroukhim, R., et al. (2010). "The landscape of somatic copy-number alteration across human cancers." Nature **463**(7283): 899-905.

Bishop (2006). Pattern Recognition And Machine Learning, Springer-Verlag New York, Inc.

Brunk, H. D. (1955). "Maximum Likelihood Estimates of Monotone Parameters." The Annals of Mathematical Statistics **26**(4): 607-616.

Carter, S., et al. (2011). "Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping." Nature Precedings.

Carter, S. L., et al. (2012). "Absolute quantification of somatic DNA alterations in human cancer." Nat Biotech **advance online publication**.

Cheung, H. W., et al. (2011). "Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer." Proceedings of the National Academy of Sciences **108**(30): 12372-12377.

Consortium, T. U. (2013). "Update on activities at the Universal Protein Resource (UniProt) in 2013." Nucleic Acids Research **41**(D1): D43-D47.

Curtis, C., et al. (2012). "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups." Nature **486**(7403): 346-352.

Dan Foresee, F. and M. T. Hagan (1997). Gauss-Newton approximation to Bayesian learning. Neural Networks,1997., International Conference on.

De, S. and F. Michor (2011). "DNA secondary structures and epigenetic determinants of cancer genome evolution." Nat Struct Mol Biol **18**(8): 950-955.

Dempster, A. P., et al. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B (Methodological) **39**(1): 1-38.

Frankild, S. and L. J. Jensen University of Copenhagen DISEASES database. http://diseases.jensenlab.org.

Kaelin, W. G. (2012). "Use and Abuse of RNAi to Study Mammalian Gene Function." Science **337**(6093): 421-422.

Marcotte, R., et al. (2012). "Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells." Cancer Discovery **2**(2): 172-189.

McLachlan, G. J. and T. Krishnan (2007). The EM Algorithm and Extensions, Wiley.

Mermel, C., et al. (2011). "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers." Genome Biology **12**(4): R41.

Morganella, S., et al. (2011). "Finding recurrent copy number alterations preserving within-sample homogeneity." Bioinformatics **27**(21): 2949-2956.

Perou, C. M., et al. (2000). "Molecular portraits of human breast tumours." Nature **406**(6797): 747-752.

Prat, A. and C. M. Perou (2011). "Deconstructing the molecular portraits of breast cancer." Molecular Oncology **5**(1): 5-23.

Ramsköld, D., et al. (2009). "An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data." PLoS Comput Biol **5**(12): e1000598.

Santarius, T., et al. (2010). "A census of amplified and overexpressed human cancer genes." Nat Rev Cancer **10**(1): 59-64.

Shao, D. D., et al. (2013). "ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens." Genome Research.

Subramanian, A., et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America **102**(43): 15545-15550.

TCGA (2012). "Comprehensive molecular portraits of human breast tumours." Nature **490**(7418): 61-70.

Walter, V., et al. (2011). "DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors." Bioinformatics **27**(5): 678-685.

Weinstein, I. B. and A. Joe (2008). "Oncogene Addiction." Cancer Research **68**(9): 3077-3080.

Weir, B. A., et al. (2007). "Characterizing the cancer genome in lung adenocarcinoma." Nature **450**(7171): 893-898.

Yuan, X., et al. (2012). "Genome-wide identification of significant aberrations in cancer genome." BMC Genomics **13**(1): 342.

Yuan, X., et al. (2012). "TAGSCNA: A Method to Identify Significant Consensus Events of Copy Number Alterations in Cancer." PLoS ONE **7**(7): e41082.

Zender, L., et al. (2006). "Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach." Cell **125**(7): 1253-1267.

Table S1. Regions of Recurrent Amplification Detected by ISAR, Related to Figure 1

Significantly amplified regions identified by ISAR in the TCGA Breast cancer cohort.

Index	Chromosome	Start	End	ISARScore
1	1	1720000	2940000	7.332
2	1	38400000	40400000	2.108
3	1	41800000	43400000	4.479
4	1	61000000	62400000	3.335
5	1	93800000	94100000	2.243
6	1	149000000	149000000	4.893
7	1	153000000	154000000	3.206
8	1	159000000	161000000	6.57
9	1	201000000	204000000	3.483
10	1	233000000	233000000	3.257
11	2	9820000	10200000	3.138
12	2	38400000	38600000	2.342
13	3	4370000	5230000	3.882
14	3	14500000	15400000	2.821
15	3	171000000	172000000	2.047
16	3	180000000	181000000	2.204
17	4	1290000	2080000	3.706
18	4	73400000	75900000	6.632
19	4	76400000	77800000	2.516
20	4	144000000	145000000	2.665
21	5	936053	1350000	2.527
22	5	13000000	14600000	4.778
23	5	44300000	44800000	3.492
24	5	176000000	177000000	2.595
25	6	41600000	42200000	2.121
26	6	43700000	44200000	2.016
27	6	63800000	64800000	2.119
28	6	105000000	109000000	19.9
29	6	135000000	136000000	2.921
30	6	151000000	152000000	4.549
31	7	5370000	5660000	2.396
32	7	23000000	23500000	2.671
33	7	32900000	33000000	2.073
34	7	55000000	56400000	4.989
35	7	68500000	68800000	2.083
36	7	98400000	100000000	5.278
37	7	156000000	157000000	4.341

38	8	9800000	10600000	2.369
39	8	37100000	39300000	15.74
40	8	81000000	82100000	5.064
41	8	101000000	103000000	4.131
42	8	116000000	118000000	5.829
43	8	128000000	129000000	14.65
44	9	33400000	36600000	5.477
45	9	127000000	130000000	5.444
46	10	61200000	63700000	3.859
47	10	76000000	76900000	3.048
48	10	80000000	81700000	3.048
49	10	123000000	124000000	6.728
50	11	19100000	19900000	2.11
51	11	32000000	35700000	4.843
52	11	68300000	70700000	54.1
53	11	76400000	78000000	8.904
54	11	118000000	118000000	3.258
55	12	416712	1000000	3.52
56	12	26700000	27100000	2.757
57	12	56300000	56600000	3.433
58	12	67000000	69300000	11.36
59	12	122000000	123000000	3.98
60	13	26300000	26900000	2.17
61	13	29100000	30200000	4.17
62	14	34300000	35400000	2.202
63	14	37000000	37400000	2.202
64	14	48900000	49900000	4.096
65	14	102000000	103000000	2.162
66	15	47900000	51000000	3.395
67	15	96000000	97600000	6.478
68	16	10300000	11800000	2.857
69	17	23600000	25100000	11.66
70	17	33700000	35800000	73.95
71	17	44300000	46900000	7.123
72	17	54500000	56900000	14.53
73	17	57300000	58700000	2.527
74	18	13400000	13700000	2.059
75	18	22100000	24000000	3.531
76	18	58200000	60100000	4.498
77	19	14800000	15500000	2.724
78	19	34400000	35300000	4.223

79	19	60200000	61000000	3.265
80	20	33100000	34400000	3.246
81	20	51000000	52600000	17.84
82	21	15600000	16300000	2.844
83	23	23600000	24400000	3.274

Table S4. Primers Used for DNA constructs, Related to Extended Experimental Procedures

Primer	Sequence
BglAGFG2-kozak-F	AAAGATCTCGCCACCATGGTGATGGCGGCGAAGAA
EcoAGFG2-R	AAGAATTCCTACAAGAAGGGGTTGGTGGTT
EcoBRF2-kozak-F	AAGAATTCCGCCACCATGCCAGGCAGAGGCCGCTGCCCGGACT
BamBRF2-R	AAGGATCCTCAGGGAGGGTTAGGGACACT
EcoC6orf203-kozak-F	AAGAATTCCGCCACCATGGCTATGGCTAGTGTTAAATTGCTT
BamC6orf203-R	AAGGATCCTTATTTAGACATTCTCTTCTTAGGCAA
EcoGNB1-kozak-F	AAGAATTCCGCCACCATGAGTGAGCTTGACCAGTTA
BamGNB1-R	AAGGATCCTTAGTTCCAGATCTTGAGGAA
XhoNIT1-kozak-F	AACTCGAGCGCCACCATGCTGGGCTTCATCACCAGGCCTCCTCACA
EcoNIT1-R	AAGAATTCTCAAGAGGAGACGGGCTCCCAGT
XhoPRKCZ-kozak-F	AACTCGAGCGCCACCATGCCCAGCAGGACCGGCCCCAAGAT
EcoPRKCZ-R	AAGAATTCTCACACCGACTCCTCGGTGGACA
EcoTRPS1-kozak-F	AAGAATTCCGCCACCATGGTCCGGAAAAAGAACCCCCCTCTGA
BamTRPS1-R	AAGGATCCCTACAGGAATCCCTTGGTTTCCA
XhoZNF652-kozak-F	AACTCGAGCGCCACCATGAGCCACACAGCCAGTTCTTGT
BamZNF652-R	AAGGATCCTTAATGATGCTGTGCTGAACT

Table S5. Forward and Reverse Sequences for Determination of RNA Levels, Related to Extended Experimental Procedures

Gene	Forward/Reverse	Sequence
C6ORF203	F	GAAGACGGGGCTAGATATTGGG
C6ORF203	R	CACTTTCACCGTTCTGCTTTTC
BEND3	F	ACTATGTGGAGGTCTACTACCCC
BEND3	R	GCTCCGGTCAAGAGACAGG
BRF2	F	GGTGGAAGACTCGCACTATTC
BRF2	R	CGACTAACTTGTTCGTTTTCCCC
YEATS4	F	GAGAATGGCCGAATTTGGGC
YEATS4	R	CCGAGCAACATTACCGTAAACT
LYZ	F	GGCCAAATGGGAGAGTGGTTA
LYZ	R	CCAGTAGCGGCTATTGATCTGAA
RSF1	F	GGATGCCGATACTATGCGTCT
RSF1	R	GCCAACTCGTTTCGATTTCTGA
PRKCZ	F	AGAGCCTCCAGTAGACGACAA
PRKCZ	R	CGGGATGAGGAAATGTAAGCAA
GNB1	F	GTGAGCTTGACCAGTTACGG
GNB1	R	TGTGATCTGAGAGAGAGTTGCAT
ZNF652	F	GCTGGTTGAAAACTGTGCTGT
ZNF652	R	GAAGATGGCACTTGACCACGA
NIT1	F	GTGTGCCAGGTAACATCGAC
NIT1	R	AGGGTCCCGTGCAATGAAG
PVRL4	F	AGGACGCAAAACTGCCCTG
PVRL4	R	TGAAGCCCGTATTTGGAGTGC
TRPS1	F	AGCCCCAGTAAGGGAGGAAA
TRPS1	R	GGGTGCAGGCCATATCTTGAG
DNAJB6	F	CATGCCTCACCCGAGGATATT
DNAJB6	R	CCTCCGCTACTTGCTTGAATTT
SETD8	F	ACCGACGGGGAGAACGTATT
SETD8	R	GCATTCCAGAGCATTTGTTCG
RPS6KB1	F	CGGGACGGCTTTTACCCAG
RPS6KB1	R	TTTCTCACAATGTTCCATGCCA
TMEM49	F	TGGCATCGTCAAAGCATTGTG
TMEM49	R	CTGAGGCTATATGTGGACCCA
RNFT1	F	CCTGAAGCAAAGACATCTGGG
RNFT1	R	ACTGTGCAGTTGGCTACGATT
CCND1	F	GCTGCGAAGTGGAAACCATC
CCND1	R	CCTCCTTCTGCACACATTTGAA
BACTIN	F	CGCAGACACCTTCTACAATGAGCTGCG
BACTIN	R	GAGGCGTACAGGGATAGCACAG