Modeling resistance to vemurafenib using integrative approaches: pitfalls and opportunities

Author: Sander Lambo

Examiner: Joep de Ligt

Second reviewer: Edwin Cuppen

16-07-2014

Contents

Abstract

Drug resistance against targeted inhibitors is a complex problem that prevents efficient treatment of many tumors. Drug resistance can occur on multiple levels, requiring the acquisition and integration of system-wide data. We review the pitfalls of data integration across different layers of biology together with possible approaches to integrate the data. Additionally, we propose a model that enables the identification of mechanisms that underlie drug resistance through the integration of different layers of biological data.

Introduction

Developing treatment strategies against various cancers using drugs that inhibit specific proteins is of great clinical interest, because conventional treatment, such as chemotherapy and irradiation of the tumor, are harmful to surrounding tissue [1]. However, drug resistance can occur within the patients through complex mechanisms that have not been fully elucidated [2]. Drug resistance can be mediated by changes in protein abundance and protein structure [3, 4]. These changes can be caused by processes on the level of DNA, RNA, protein and metabolism [3, 4]. Therefore, investigating mechanisms of drug resistance requires the acquisition of data at a large scale and on different biological levels. Recent developments enabled the application of large scale data acquisition overcoming some of the challenges in creating complex models. For instance, modeling of a complete cellular system in mycoplasma using many large scale datasets allowed the prediction of phenotype from genotype [5]. To create this model genomic, transcriptomic, proteomic, metabolic and other factors such as mass and time were unified in a single model. However, unifying different layers of data in a system-wide manner, also known as integration, remains complex, especially for multicellular organisms with large genomes [6]. In this review potential integration approaches are evaluated for their applicability in identifying and predicting resistance to targeted inhibitors. These approaches will integrate genomics, transcriptomics and proteomics data to identify pathways involved in drug resistance but will be limited by our understanding of biology and the currently available methods to acquire data.

Resistance to targeted treatment

Vemurafenib is a drug commonly used in melanoma treatment. Vemurafenib targets mutant BRAF[*] which is mutated at position 600 (BRAF(V600E)). BRAF is part of the MAPK[†] pathway that is involved in processes such as survival and proliferation. An activating BRAF mutation occurs in 50 to 70% of all melanomas and 90% of these mutations constitute BRAF(V600E) mutants [7]. The MAPK pathway consists of receptor tyrosine kinases (RTKs) that activate NRAS[‡]. NRAS activates BRAF, which phosphorylates MEK[§] leading to the phosphorylation of ERK[**] [8].

Therapies using vemurafenib are highly effective in melanomas but have shown little effect combined with quick relapse in BRAF(V600E) colon tumors [9]. Although treatment is initially highly effective in melanoma the duration of the effective treatment is six months on average [9]. Resistant tumors expand and invariably lead to a fatal outcome [10]. Recently, combination treatment has been identified as a viable strategy to extend the lifespan of patients [11]. Unfortunately, this strategy was not effective for all tumors as drug resistance can occur through several mechanisms involving multiple pathways. As a result, combination treatment can be ineffective in a subpopulation of patients. Lack of treatment response in patients requires the detection of mechanisms involved in drug resistance and adaption of treatment on a patient to patient basis (personalized treatment).

Personalized treatment requires the understanding of mechanisms involved in drug resistance. Drug resistance can occur through reactivation of the inhibited pathway or activation of an alternative

[*] v-raf murine sarcoma viral oncogene homolog B

[†] mitogen-activated protein kinase

[‡] neuroblastoma RAS viral (v-ras) oncogene homolog

[§] mitogen-activated protein kinase kinase

[**] extracellular-signal-regulated kinase

pathway [12] (**Figure 1**). For instance, activating mutations in NRAS can reactivate the MAPK pathway, but also upregulation of C-MET and downstream factors were found to confer resistance [13, 14]. This "rerouting" of pathways eventually stimulates proliferation and survival capacities of the cell. It has been proposed that cells with increased proliferation rate and survival capacity have a growth advantage over cells that remain affected by the drug, resulting in exponential expansion of resistant cells [2]. However, there is some debate on whether alterations stimulating growth advantage are present in a subpopulation before treatment or if cells acquire alterations upon or after treatment. For instance, Strausmann *et al.* showed that a resistant subclone is formed by a subpopulation with higher abundance of C-MET receptors before treatment, called innate resistance [15]. Conversely, Sun *et al.* showed that expression of EGFR, normally not expressed in melanoma cells, was acquired in drug resistant subclones, called acquired resistance [2].

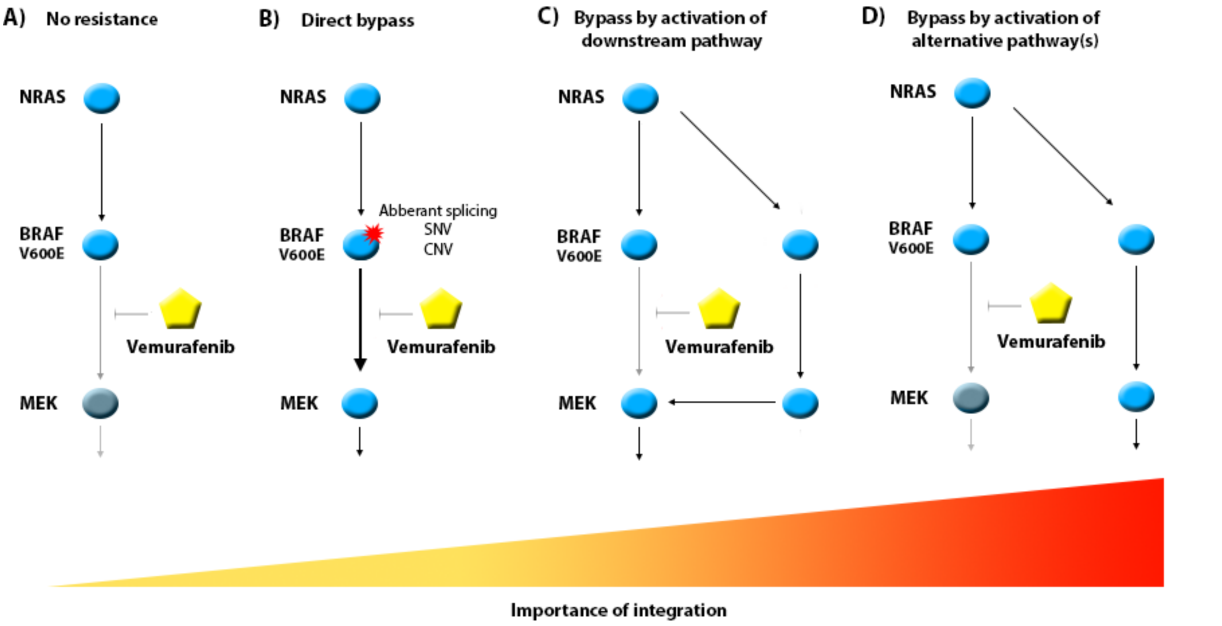
A) No resistance
B) Direct bypass
C) Bypass by activation of downstream pathway
D) Bypass by activation of alternative pathway(s)
NRAS
BRAF V600E
MEK
Vemurafenib
Abberant splicing
SNV
CNV
Importance of integration

Figure 1. Mechanisms underlying resistance to vemurafenib

Schematic depiction of the mechanisms underlying drug resistance to vemurafenib. **A)** Cells respond to treatment: MEK is not activated by BRAF. Bypassing inhibition by vemurafenib leads to disease progression (B-D). **B)** Drug resistance can occur by increased amounts of active BRAF (CNV) or by limited effectiveness of the drug (SNV or aberrant splicing) leading to more active MEK (larger arrow). **C)** The MAPK pathway can be activated downstream through upregulation of proteins that also regulate downstream factors such as MEK. **D)** Resistance can occur through upregulation of alternate pathway(s) that exhibit the same function as the inhibited pathway.

Identification of the alterations that cause the reactivation of pathways remains difficult. One major complicating factor is tumor heterogeneity, meaning that not every cell within the tumor contains the same alterations. By gathering data from entire tumors, heterogeneity may cause alterations within a subgroup of cells to be missed because phenotypes are averaged over the entire tumor [16]. Additionally, malignant cells are thought to contain an instable genome with a higher rate of mutations compared to healthy cells, which can lead to a large number of mutations [17]. Many mutations will not contribute to drug resistance (passenger mutations) compared to a few mutations that do (driver mutations). The increase in the total amount of mutations impedes the detection of driver mutations, since the amount of noise caused by passenger mutations is increased [2].

Cancer is a disease caused by mutations in the DNA, but it has been shown that resistance to targeted treatment can be caused by alterations on other levels as well [15, 18]. For example, alternative splicing of BRAF mRNA can confer drug resistance [18]. Increased C-MET signaling can occur through higher secretion of HGF followed by upregulation of proteins in the same pathway, an example of combined metabolic and protein level changes leading to drug resistance [15]. Using data across different levels, known as omics, is necessary to detect mechanisms behind drug resistance in individual patients.

Integration of omics data

Resistance occurring at multiple levels signifies the need of obtaining different types of omics data. Curtis *et al*. described the association of single nucleotide variation (SNV) and structural variation (SV) with expression changes. This study resulted in novel associations between SNVs and expression changes of genes involved in the immune response in addition to known oncogenes. Furthermore, the increase in expression of the identified genes was found to be part of the same network by using online databases such as KEGG, BioCarta, PANTHER and cancer cell map [19]. Balbin *et al.* described a method to integrate gene expression, protein abundance and abundance of phosphorylated proteins to reconstruct overexpressed networks in lung cancer. By using this approach three different networks were found [20].

Curtis *et al*. identified a network of functional categories, such as immune response, and the underlying cause in the DNA, but could not find the individual proteins driving the resistance. In addition, the network was based on existing databases containing interaction networks in healthy tissue where there is no rerouting of pathways [19]. Balbin *et al.* could distinguish which protein was important in the network but lacked information about the cause of the malignancy. This method also resulted in three networks with related subnetworks but did not distinguish which network was more significant in causing the phenotype [20].

The study by Curtis *et al*. integrated genomic and transcriptomic data in breast cancer while the study by Balbin *et al.* integrated transcriptomic, proteomic and phopshoproteomic data in lung cancer [19, 20]. Both studies identified pathways that drive malignancy. However, the difference between the studies is that one study identified upregulated pathways whilst the other identified upregulated parts of pathways. Obtaining information about differentially regulated pathways is important but acquiring additional information about which protein to target is preferred over solely identifying the pathway that drives resistance. Therefore, finding the cause and significant pathways involved in resistance requires an integrative approach of genomic, transcriptomic and proteomic data.

From genotype to phenotype: pitfalls in integration

Integrating data using a linear dogma of transcription and translation (DNA -> RNA -> protein) that influences the phenotype has led to the discovery of many genotype-phenotype relations [21-23]. However, this linear dogma cannot be applied to every genotype-phenotype relation since the effect of mutations in DNA can be small in the eventual protein since not every mutation will result in structural changes in the protein [24]. Additionally, events on the RNA level can affect protein structure. Changes in protein structure can affect protein stability and interactions between proteins possibly leading to an altered phenotype [25].

Integration of data is not as straightforward as the central dogma suggests. For instance, the association between variation in DNA and gene expression was found to be approximately 40% [19]. The correlation between RNA expression and protein abundance was also found to be relatively low (40-50%) [6, 26]. Processes that hamper integration occur at DNA level, RNA level and protein level, eventually leading to changes at protein level. Events can be broadly classified in two types: processes that eventually influence structure and function of proteins and processes that eventually influence the abundance of proteins.

Processes involved in altering protein structure

Tumor cells contain many single base pair deletions, insertions (InDels) and substitutions (SNV). These changes can alter protein structure and change either protein stability or interactions with other proteins. Other forms of variation can influence large stretches of the genome (SVs), such as entire chromosome arms, by causing translocations, inversions and deletions. Both SNVs and SVs can result in no change in amino acid (synonymous), a substitution of amino acids (missense), or termination of translation (nonsense). Synonymous variations generally have a less direct effect on the eventual protein structure compared to missense mutations while nonsense mutations often lead to non-functional proteins [24]. It is possible to predict the effect of missense mutations on protein structure to some extend by using tools such as SIFT and PolyPhen2 [24, 27, 28].

Predictive tools can estimate the eventual effect of mutations but do not correct for processes occurring at RNA level. For instance, bases in the RNA can be edited by an editosome leading to a different amino acid sequence [29]. RNA editing is relatively rare, but can lead to structural changes by introducing more SNVs. Other, more common, events such as alternative splicing can also influence protein structure. Splicing is a process that mainly cuts out introns, but also exons, from a coding region. Alternative splicing can lead to translation of a protein with different functionality or degradation of RNA molecules [30]. Another event that influences the effect of mutations is caused by the presence of multiple copies of the same gene. The mutations in the separate copies can cause a bias in transcription leading to mutant alleles being overexpressed [31].

Detecting events such as SNVs, SVs, RNA editing, splicing and allelic biases from DNA and RNA sequencing data is feasible. A major limitation is predicting how these changes affect protein interaction, stability and function. This can be predicted using homology modeling and fold prediction, however the accuracy of these predictions is low [32, 33]. Besides lack of accuracy, protein stability and interaction after structural changes cannot be robustly predicted because of the low amount of available crystal structures and docking studies to validate predicted changes [34]. Finally, there are post-translational modifications (PTMs), such as phosphorylation and ubiquitination, which can affect the structure, activity, localization and interactions of proteins [35]. Changes in phosphorylation levels are relevant for modeling drug resistance, since the MAPK pathway is a kinase cascade and the activity of the pathway can be estimated through phosphorylation levels.

Processes involved in altering protein abundance

Processes influencing total protein abundance occur on multiple levels similarly to events influencing protein structure. On the genomic level there are genome duplication events that cause genes to be more abundant and generally lead to higher expression. Genes can also be lost by loss of chromosome arms during cell division. These changes in gene copies are often named copy number variants (CNVs) [36]. Next to the amount of copies of a gene there are DNA sequences influencing transcription. Variations outside coding regions such as enhancer elements influence the expression of genes. The understanding of transcriptional regulation by enhancer elements is very limited. Recently, techniques such as 4C and 5C were developed to obtain data on genomic structure in order to investigate which enhancer influences which gene [37]. However, most physical interactions have not been mapped at this time [37]. Other factors complicating the determination of transcription efficiency are both epigenetic marks and chromatin on the DNA. To investigate the effect of both epigenetics and chromatin techniques such as bisulfite sequencing and DNAse sequencing need to be used. However, even with the addition of this data, our knowledge of both chromatin and epigenetic marks on expression is limited [38, 39].

Several other processes influence mRNA abundance after transcription; microRNAs (miRNA) can decrease the expression of genes by binding to the mRNA molecule. miRNAs have been found to be informative in classifying malignancy in tumors and may play a significant role in drug resistance as well [40]. In addition to the efficiency in transcription, turnover of mRNA named nonsense-mediated decay (NMD) can lower the abundance of proteins. NMD can be caused by nonsense mutations, structural variation or premature splicing [41].

Metabolic processes also play a role in protein abundance. Degradation and translocation of proteins lowers the amount of active proteins in a network, eventually influencing the phenotype. Metabolomics is modeled using the quantities of metabolites in the system [42]. However, the amount of metabolites changes significantly over time complicating the prediction of phenotypes. One way to model this is using flux models that predict the variability of the amount of metabolites in the system [43].

The influence of amino acid substitutions on the eventual protein function and the quantification of the amount of DNA that is transcribed into RNA are significant pitfalls in integration (**Figure 2**). However, all cellular processes occur over time and omics data is usually taken from one transient state [44]. This is a substantial pitfall as acquiring large datasets over longer periods of time is very costly. In addition, it is often not feasible to acquire the necessary amounts of (tumor) material.

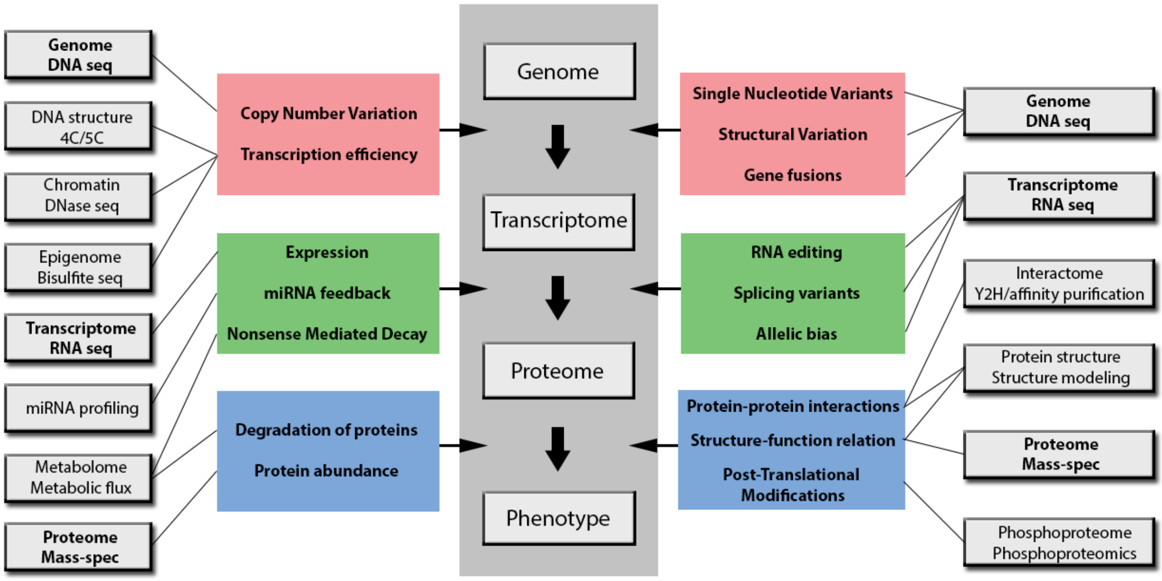

Figure 2. Biological complexity influencing data integration
Novel insights continue to increase the complexity of the central dogma (gray backdrop). Processes and variation influencing the abundance of proteins are shown on the left, those influencing protein structure (and subsequent interactions) are shown on the right. Red boxes are processes occurring on DNA level, green boxes are processes occurring on RNA level and blue boxes are processes occurring on protein level. On the far left and right side are different omics data types that influence the central dogma with the proposed technique to acquire the data. Techniques shown in bold are integrated in the model.

Creation of an integrated model

Modeling the biological complexity of drug resistance requires the development of an integration method for processes involved in conferring resistance. Modeling the entire complexity is not feasible and necessary at this time, therefore a selection of processes that is possible to integrate using current available data should be made. Predominantly, the approach should detect pertubated networks that cause resistance, for instance upregulation of C-MET and downstream factors [14, 15].

Whole cell models such as the model presented for mycoplasma divide processes into parts (modules) to create structure [5]. These modules can be integrated to form a weighted module that predicts the perturbation and eventually the phenotype [5, 45]. As shown by Zhang *et al*. a matrix of copy number variation and somatic mutations at every position can be integrated into a module based on multiple modules (cluster) of DNA level changes [45]. By using a matrix of expression data a cluster was created at the RNA level. The DNA level and RNA level clusters were integrated into a single network [45]. By using this method a subnetwork was identified, independent of known protein interactions, which could potentially drive malignancy in glioblastoma multiforme (GBM) [45]. The identified pathway was involved in regulation of the cell cycle, which is important in proliferation and could lead to a growth advantage [45]. Using more modules to infer a more accurate network is possible: in mycoplasma 28 modules were integrated to accurately predict a phenotype [5].

Networks

Modules need to be connected in a network to enable the integration of data across different layers. By connecting modules subnetworks, relevant in causing drug resistance, can be detected and defined. There are several methods to combine modules into networks of which the five most relevant models are briefly discussed here:

> I Seed- and extend methods detect relevant subnetworks by the density of interactions between data. For instance, proteins that act in the same complex have many shared interactions. Therefore this method is usually deployed to detect protein complexes [46].
>
> II Frequency-based methods utilize known data to find subnetworks that are present in multiple datasets. For instance, subnetworks around p53 are found to be important for many cancer types. These methods are usually deployed to find common pathways in a large set of samples [47].
>
> III Hierarchical clustering is based on the similarity of two datapoints. For instance, genes that exhibit the same expression changes under the same condition cluster together. Similarly, proteins can be clustered based on the similarity of interaction partners [48]. This method is more relevant for our question but can only detect pathways based on one layer of data at a time. Therefore, this method is not the best choice for data integration from different layers of omics data.
>
> IV Optimization based methods use algorithms to reconstruct pathways based on significantly changing datapoints and attempts to reconstruct networks which use the same amount or less connections as networks found in the background population [49]. This method was optimized for creation of subnetworks based on the influence of mutations. However, this method requires complex algorithms that are not fully developed and tested at this point [50].
>
> V Statistical methods use probabilities to test which subnetwork has a high chance of being regulated differently from the rest of the network. An example is Bayesian inference, a statistical method to update probabilities after addition of evidence (data) [51]. This can be used to form a Bayesian network of factors influencing the final probability [52]. In resistance modeling the probability that a subnetwork is driving the resistance is the intended result and the acquired omics data the evidence.

Based on these characteristics a Bayesian network is the most likely choice to integrate the data and to find subnetworks driving resistance. Finding subnetworks that drive resistance by using only DNA mutational data is possible as was shown by a program named VarWalker. Yet, the consensus pathways that were reached contained a very broad spectrum of mutations and it was impossible to distinguish the driver mutations from the passenger mutations [53]. The paper by Zhang *et al.*, which used copy number variation and expression data in addition to mutational data, reached a smaller number of consensus pathways which were more accurate [45]. Therefore, modules can be added to decrease the rate of false positives in the final prediction [52].

Addition of modules can alter the probabilities of clusters as well. For instance, the existence of a perfect correlation between RNA levels and protein abundance for a specific gene decreases the likelihood that miRNA signatures are an important process in driving the resistance [40, 52]. Bayesian networks allow for integration of multiple types of data, correction of the interaction between modules and prediction of functional pathways. Therefore, Bayesian inference is the method of choice to integrate the data and detect resistance mechanisms.

Modules to integrate changes in protein structure

Mutations can have different consequences on the eventual protein structure (see "Processes involved in altering protein structure") and can be used to model the probability that structural changes play a role in conferring resistance [24]. Variations such as InDels can cause translation to occur in another reading frame (frameshift mutation) that causes changes in all consecutive amino acids. As such, frameshift mutations have a larger effect on the structure compared to substitutions [54]. SVs usually underlie a deregulation in large stretches of a protein and can result in truncated proteins [55]. Together these variations can be used to assign a severity score to mutations since mutations with a more pronounced effect will be more important in protein structure, stability and interactions.

Structural changes caused by alterations on RNA level can be detected fairly well due to recent advances in RNA sequencing technology [56]. Discrepancies between DNA and RNA data can be viewed as RNA editing events, although errors in transcription by RNA polymerase can also cause changes in sequence [29]. Other RNA events require more effort to detect. For instance, allelic bias can be examined by ligating adapters to the primers on each orientation. This allows for the detection of increased expression of one allele [57]. Splicing variants can be identified by tools, such as DEXseq, that use statistical modeling to predict the occurrence of splice events based on sequencing data [58]. Alternatively, reads can be mapped to the spliced exome, enabling detection of splice events since spliced out exons will have a decreased number of reads. Splicing and SVs can cause genes to be transcribed directly after other genes leading to a protein that has an amino acid sequence derived from two genes. This is also known as gene fusion and can be detected using the same methods as aberrant splicing. Together, events on RNA level can be qualified similarly to DNA events.

Creating a structure from an amino acid sequence remains a major pitfall in integrating structural variation. Possible approaches to classify amino acid substitutions are based on position; amino acid substitutions in the proximity of an interaction domain or phosphorylation site generally have a larger effect [34, 59]. Taken together, currently the only possible way to integrate processes involved in protein structure is by classifying predictions on protein function alterations (**Table 1**) [25].

Table 1. Modules involved in protein structure

Modules are divided in categories: Quantitative (based on the amount of events) and Qualitative (based on classification). Modules greyed out are not included in the model but can eventually be added.

Module	Type	Probability based on
SNV	Qualitative	Position/effect aa substitution on structure and interaction
SV	Qualitative	Gene fusions / Truncations / InDels
Frameshift	Qualitative	Amount of affected codons/ early stop codons
Allelic bias	Qualitative	Strength bias towards mutated allele
Aberrant splicing	Qualitative	Physicochemistry/function alternative protein
RNA editing	Qualitative	Position/effect aa substitution on structure and interaction
Phosphorylation	Quantitative	Increase/ decrease phosphorylation levels
Structural change	Qualitative	Folding / Physicochemistry / Accessibility functional sites
Interaction change	Qualitative	Changes in interaction partners

Modules to integrate changes in protein abundance

Integration of alterations on the DNA level and expression is a major challenge in integration. For instance, epigenetic marks and chromatin have been shown to substantially influence expression [38, 39]. In addition, most of the variation underlying expression changes lie in noncoding regions. The effect of noncoding events is mostly unknown as our understanding on this type of variation is limited [60]. One approach to predict which variant underlies expression changes is using expression quantitative trait loci (eQTL). eQTL studies associate noncoding variance with expression changes in nearby genes. By using known eQTL studies such as the study performed by Westra *et al.* variance within noncoding regions can be associated with gene expression [61]. Variants associated with a disease can be positively scored in the model. CNVs, however, correlate well with expression changes as was shown before in breast cancer cell lines. In cell lines a positive correlation of 64% was found between copy number gain and expression [62]. Therefore, copy number gain or loss can be modeled by adding positive scores for copy number gains and negative scores for copy number losses.

Gene expression data obtained from the RNA sequencing data can be scored using the same method as CNVs [63]. Yet, CNVs have a low correlation with the actual protein abundance [64]. This could be largely explained by the weak correlation between expression and protein abundance [6, 26]. Finally, the observed protein abundance using proteomics can be put in another module comparing the observed abundance with a wild-type situation and giving positive scores to increases and negative scores to decreases. The different modules affecting abundance and their effect on the abundance are shown in **Table 2**.

Table 2. Modules involved in protein abundance

Modules are divided in categories: Quantitative (based on the amount of events), Qualitative (based on classification) and association (based on effect on other level). Since both increase and decrease in abundance can lead to progression the events that lead to either an increase or decrease in abundance are shown. Modules greyed out are not included in the model but can eventually be added.

Module	Type	Protein abundance increase	Protein abundance decrease
CNV	Quantitative	Copy number gain	Copy number loss
eQTL	Association	Increased expression	Decreased expression
Epigenetics	Qualitative	Stimulating marks	Repressive marks
Chromatin	Qualitative	TF binding sites accessible	TF binding sites inaccessible
Expression	Quantitative	Increased mRNA	Decreased mRNA
miRNA profiles	Quantitative	Decreased presence signature	Increased presence signature
NMD	Quantitative	Decreased turnover	Increase turnover
Metabolic flux	Quantitative	Decreased turnover	Increase turnover
Abundance	Quantitative	Increase protein levels	Decreased protein levels

Protein interaction networks and genetic interaction networks

Different modules were formulated as probabilities influencing resistance. Subsequently, these modules need to be integrated in a model that can predict pathways that drive resistance.

Omics data is commonly used to infer protein interaction networks and genetic interaction networks [24]. Protein interaction networks are generally used to predict the function of a protein based on protein interaction. Protein interaction networks are based on ‘guilt-by-association’: a protein with an unknown function interacts with other proteins creating a complex or pathway that provides clues about the function of the protein of interest. However, a network using protein-protein interactions only provides information on which proteins interact but what kind of interaction is unknown in such a network [65].

Genetic interaction networks are generally used to predict function and interaction between genes (epistasis) [24]. Genetic interaction networks depict whether genes increase or decrease the expression of a gene and can be used to infer pathways. However, genetic networks commonly depict mutations as loss of function mutations only resulting in genes being removed from the interaction network. It is possible that gain of function mutations or changes in function occur, melanoma being no exception [66]. Therefore, mutations can change the eventual network by inhibiting other proteins, increased stimulation of other proteins or by different functionality.

Genetic interaction networks are traditionally used to connect sequence variation with function [24]. However, we need to connect changes in all layers with function, since it is possible that not all resistance mechanisms have underlying mutations. Therefore, a model will be proposed based on modules influencing protein abundance and protein function (**Figure 3**).

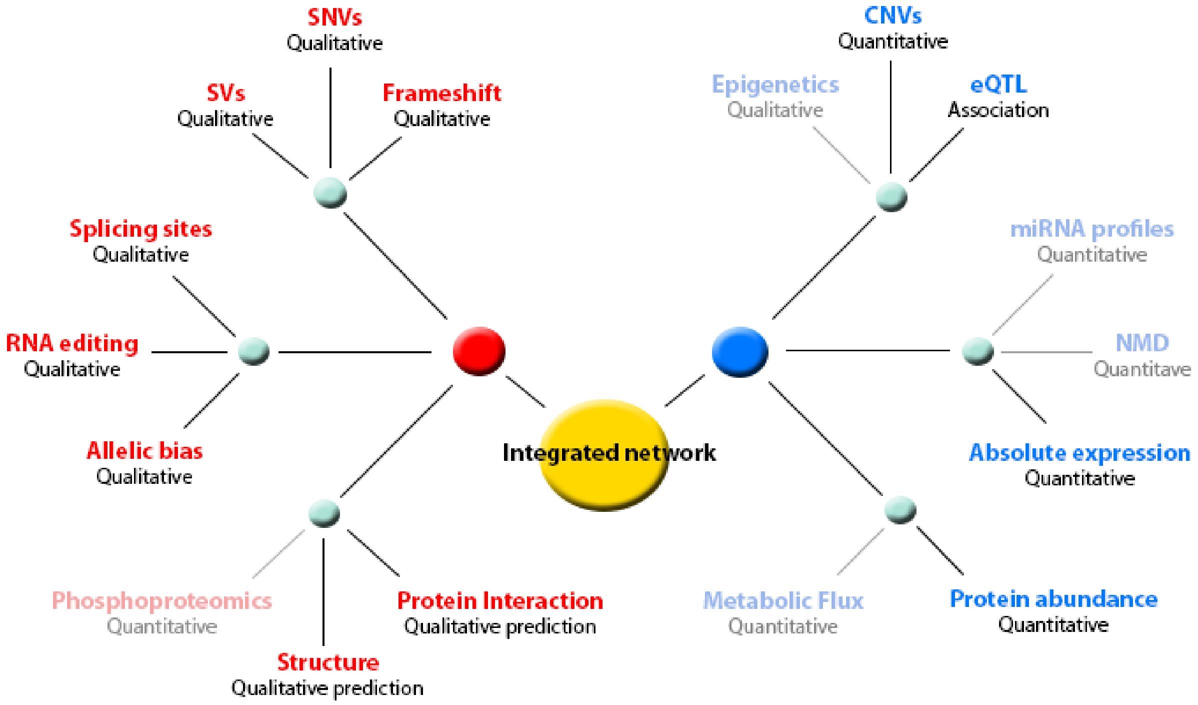

Figure 3. Integration based on protein abundance and protein structure

Integration using six clusters derived from processes that either influence protein structure or protein abundance at different levels. Nodes (cyan) represent clusters influenced by the separate modules or other clusters. Influences are shown as lines (edges) between the nodes. Modules influencing protein structure are shown in red and modules influencing protein abundance are shown in blue. Modules that can be integrated by addition of other techniques (**Figure 2**) are transparent.

The model based on protein abundance and protein structure has modules representing probabilities that proteins are differentially regulated. The integrated network is created based on the assumption that the resulting pathways have multiple proteins that are differentially regulated. The integrated network eliminates the need to treat structure and abundance data differently since positive influences on interactions are weighted as heavy as negative influences on interactions. The afore mentioned model integrates data based on clusters derived from processes at omics level. These six clusters are integrated to infer two integrated networks that influence either protein abundance or protein structure, and are subsequently used to detect processes that drive resistance.

Inferring networks from modules

The model shown in **Figure 3** assumes that the contribution of every module is equal. However, in practice modules such as protein abundance have a larger effect on the eventual network compared to for instance the presence of CNVs. Normally, model training has to be performed to apply weights to input modules but can also be inferred using Bayesian methods [67]. Bayesian methods have three functions: inferring unknown variables, parameter learning and structure learning [68, 69]. Inferring unknown variables in resistance modeling consists of determining differentially regulated protein abundance or protein structure based on changes on DNA, RNA and protein levels. This can be performed directly from the probabilities found in the modules. Parameter learning is a method to estimate missing values in a dataset. For instance, when taking expression data over time some timepoints may be missing. In the simplest form these values are calculated based on the mean and variance of the rest of the data. The use of parameter estimation can be limited by generating data for all layers of integration and therefore the reliability can also be improved [69, 70]. Structure learning is less feasible to limit without using computation.

Modeling interactions between modules

Structure learning is a machine learning method to apply weights and direction to interactions between modules. Processes influencing abundance (shown in blue in **Figure 3**) are related, for instance methylation marks were shown to mark active promoters and likely increase gene expression [71]. Similar relations exist for processes involved in structural changes (shown in red in **Figure 3**), for instance mutations can cause aberrant splicing by forming a splice site [72]. In addition, there are processes that influence abundance affecting structure and vice versa, for instance aberrant splicing causing NMD [73].

Applying weights and direction between modules in a model is difficult to perform without complex algorithms since the interactions between modules are generally unknown. For instance, the correlation between expression and protein abundance is about 40-50% [26]. This correlation between expression and protein abundance can differ between classes of proteins further complicating interactions. For example, Low *et al*. found that proteins involved in the cytoskeleton had a better correlation between expression and protein abundance than membrane proteins [6].

Applying correlations between modules can be performed by using weight matrixes. Kim *et al*. described a method to integrate the interaction of CNVs, miRNA profiles, methylation profiles and gene expression data. The integration of two modules (similar as in **Figure 3**) was supplemented with a weight matrix containing the correlation between the two data types creating a submodule between modules influencing expression. This model resulted in a more accurate prediction than an integrated model which used every layer separately [74]. However, these weight matrixes were based on known data from the TCGA atlas and cannot be applied directly in a personalized setting [74].

Integrating data from the 12 modules described before requires the knowledge of interactions between data. However, a single SNV can have an effect on multiple modules (**Figure 4**). It is possible to annotate the correlations between data (weights) by hand and determine the interaction based on literature. Drawing an interaction network without computation is difficult since the direction of interactions, weighting of interactions and indirect interactions need to be applied. The structure of Bayesian networks is traditionally shown in directed acyclic graphs (DAGs). DAGs are networks that show interactions between modules in a single direction. At the end of the network lies a single module that is the intended result. In resistance modeling the final module would be the probability that a protein is involved in resistance and the modules underlying it would be the omics data [75, 76].

Using machine learning algorithms such as Markov Chain Monte Carlo (MCMC) is another possibility to apply interactions to a model. MCMC approaches create integrals based on modules that exist in a multi-dimensional space. The distance between modules can be calculated with "walkers", abstract entities that move randomly through the multi-dimensional space. Every step the "walker" makes is scored and saved until it hits another module. The closer the modules are the less steps it costs and the more the two modules are related [77].

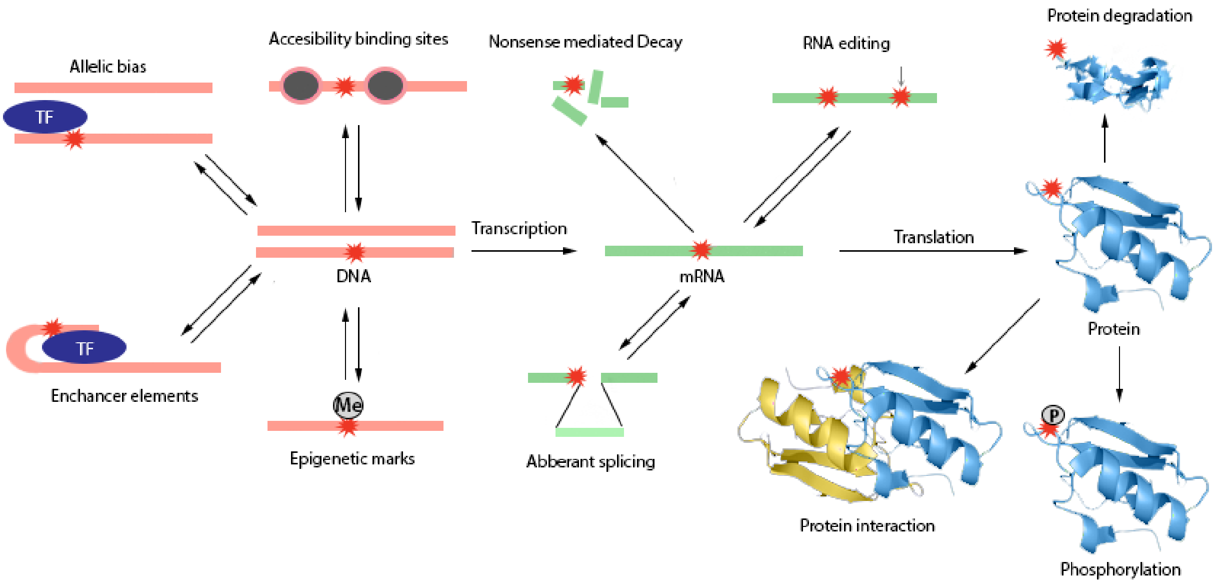

Figure 4. Biological complexity resulting from SNVs
An SNV (shown as a red star) can influence multiple processes that have an effect on transcription, protein interaction and stability. Arrows show interactions between processes, double arrows show feedback mechanisms. DNA is shown in red, RNA is shown in green, proteins are shown in blue, chromatin is shown in gray, transcription machinery is shown as "TF" and protein interaction partners are shown in yellow.

Discussion of the model

The proposed model uses a Bayesian approach to integrate three layers of biological data formulated as matrixes of probabilities to identify subnetworks that potentially drive resistance. It is likely that this approach is able to identify proteins involved in resistance since this was shown with a similar approach which integrated three modules [45]. The question is whether this method will be sufficient to reliably identify all the resistance mechanisms active after treatment.

In **Figure 3** the modules influencing protein structure and protein abundance are shown. Most of the processes influencing structure can be integrated by using only DNA and RNA sequencing techniques to generate data. However, the data is qualitative since the eventual interaction after amino acid substitution cannot be quantified. Most processes influencing abundance can be measured quantitatively except for processes underlying transcription regulation that are not well understood. Therefore, achieving a fully integrated model requires structural changes to be quantified and improved understanding of transcription regulation.

Using a fully integrated approach to find the driving cause of drug resistance may not be necessary. The majority of validated resistance mechanisms is based on upregulation of proteins in either the MAPK pathway or an alternative pathway and can be detected using quantitative measurements [2]. Identifying underlying mutations is not as important as protein abundance to define which protein to target in an upregulated pathway using targeted treatment. Therefore, the lack of understanding on integration of transcriptional efficiency will probably not decrease the feasibility of predicting relevant subnetworks. Detection of structural variants, such as an aberrant spliced form of BRAF, could be of greater importance since there is a possibility that novel splice forms of BRAF or downstream factors exist [18]. Conversely, secondary mutations in BRAF that prevent inhibition by vemurafenib were found in cell lines but not in patients indicating a smaller role of mutations in causing a resistant phenotype [2, 78].

Lack of phosphoproteomics data is likely a strong limiting factor at this point since one of the general mechanisms of resistance is downstream reactivation of the MAPK pathway. BRAF and downstream targets, MEK and ERK are kinases suggesting that phosphorylation status is more significant than abundance [8]. Vemurafenib prevents the phosphorylation of MEK and therefore does not result in lower abundance of MEK. However, vemurafenib does lead to lower abundance of phosphorylated MEK (p-MEK) illustrating that phosphoproteomics is an important layer in this network [79].

Relief of feedback inhibition is a process involved in drug resistance by causing an increase of upstream protein abundance. For instance, p-MEK or p-ERK normally inhibit NRAS implying that inhibition of p-MEK or p-ERK results in stimulated NRAS [80]. Eventually, this can accelerate drug resistance by upregulation of pathways that bypass BRAF. Relief of feedback inhibition can only be detected using a combination of proteomics and phosphoproteomics data since both phosphorylation and protein abundance are involved.

Future improvements of the model

Several techniques were considered that could be integrated to improve the chance of identifying pathways and causes underlying drug resistance. In the model five extra modules (transparent in **Figure 3**) can be included in the network, provided that the data is available. Phosphoproteomics has been discussed in the previous paragraph and can be integrated by quantifying the amount of phosphorylated proteins.

Modules such as chromatin organization and epigenetic marks are harder to integrate than phosphoproteomics data. The function of many epigenetic marks and DNA elements on transcription regulation are unknown. The ENCODE project was founded to identify DNA elements involved in transcription regulation but is not complete yet [81]. The ENCODE project found both stimulating and inhibiting methyl marks and chromatin structure at protein binding sites to be associated with increased expression [71, 81]. However, there may be other non-coding DNA elements involved in transcription regulation since the majority of DNA is non-coding [60].

Metabolic processes and miRNA signatures can be integrated fairly well to determine RNA abundance. miRNA signatures can be detected by using bead-based flow cytometric miRNA expression profiling to quantify the effect of miRNA abundance. These signatures can be quantified in different types of cancer and was also performed for melanoma [40]. Other processes involved in reducing abundance can be profiled using metabolomics. The turnover of RNA molecules by NMD and turnover of proteins can be measured and integrated in the model to explain the lack in correlation between RNA and protein level [82].

Time

One factor that cannot be modeled with the approach described here is how protein interactions vary over time. First of all there is a difference in the speed at which rewiring of pathways occurs in colon cancer cells compared to melanoma cells [9]. Detecting the mechanism behind this difference could be facilitated by a model based on protein production over time. For instance, a study by Nakakuki *et al.* gave various insights in mechanisms occurring transiently in resistance and processes occurring over longer periods of time [44]. They developed an approach that calculated the amount of *c-fos* expression, which is transcribed as a result of the MAPK pathway. Stimulated by epidermal growth factor (EGF) signalling the MAPK pathway affected *c-fos* expression transiently while Heregulin (HRG) signalling lead to sustained *c-fos* expression [44].

Measuring or modeling the amount of metabolites over time (flux models) is another way to create a dynamic model [83, 84]. For instance, flux models can be used to model the amount of proteins or phosphorylation in a pathway. This was shown by the study of Lerman *et al*. which integrated metabolic flux and gene expression in a single model [85]. Integration of both expression and metabolic flux resulted in additional information on transcription efficiency and translation efficiency over time compared to modelling only metabolites. Additionally, the flux models could be used for prediction of growth rates that could facilitate the detection of subcolonies with a growth advantage. However, prediction of quantitative changes in a dynamic system showed significant discrepancies between the *in silico* analysis and *in vivo* data [85].

Models such as described for *c-fos* expression illustrate that observed changes in expression or abundance can be transient. Nevertheless, applying such a model to multiple proteins requires more knowledge about protein dynamics [44]. Modelling the amount of metabolites using flux models has been well described for steady-state conditions, however tumors have a highly dynamic growth environment caused by the heterogeneity of a tumor and differences in supply of nutrients [83, 86]. Applying flux models to a dynamic system lowers the accuracy of predicting variations in metabolites [85]. Therefore, detecting the sustainability of rewired pathways in a dynamic system is one of the major limitations in creating models to predict resistance.

Predicting clonal expansion

The tumor is a heterogeneous structure with different mechanisms by which cells survive and proliferate. Cells within the tumor have different mechanisms by which drug resistance occurs as well. The difference in resistance mechanisms between cells makes prediction of cells that clonally expand and become resistant of great clinical interest. However, these predictions have not been extensively studied in tumors.

The prediction of drug resistance in human immunodeficiency virus (HIV) infection has been studied more extensively than in melanoma. The genotype of HIV can be determined and used to predict resistant phenotypes in the cell using linear regression models [87]. This has been performed for multiple types of drugs and integrative approaches are being developed that integrate genotype-phenotype relations together with data from clinical output to develop treatment regimens [87, 88]. This approach of predicting drug sensitivity could also be beneficial for cancer patients. However, resistance to BRAF inhibition occurs on multiple levels and therefore sequencing only the DNA will not suffice as a biomarker. It is possible to discover biomarkers with an integrative approach such as alternative splicing and upregulation of proteins [15, 18]. Perhaps this will open more possibilities to predict which subclones might expand.

Prediction of future alterations that cause cells to clonally expand is another method to predict drug sensitivity. For instance, evolutionary branching was found in tumors by multi-region sequencing. Evolutionary branching revealed that 69% of the mutations driving resistance did not occur throughout the entire tumor. Therefore, it was proposed that common branches in tumor should be used as biomarkers instead of mutations [16]. However, the effects of mutations in DNA seem to be limited in conferring resistance [2]. In addition, another paper described that Darwinian evolution does not play a substantial role in resistance but rather overexpression of proteins based on Lamarckian instruction drives resistance [89]. Lamarckian instruction is a model commonly used in predicting stem cell states and is based on cells expressing proteins "instructing" other cells to express the same proteins [89]. However, to observe processes such as Lamarckian instruction it is necessary to obtain proteomic data at single cell level, which is difficult to perform for entire subclones in the tumor [90].

Obtaining tumor material

Increasing the amount of tumor material in order to integrate time, more layers of omics, and spatial effects is an important step to improve the model. An obvious approach to increase the amount of tumor material is expanding the cells in a homogeneous culture. However, this approach cannot reproduce the heterogeneity in the tumor, which will lead to mechanisms such as increased HGF secretion by stroma to be undetectable [15]. In addition, the structure of the tumor is not reproduced which will lead to an equal distribution of growth factors and nutrients to the colonies. This change in distribution can cause the silencing of tumor suppressor genes such as PTEN and changes in clonal outgrowth [91]. Perhaps a 3D culturing method such as organoid culturing supplemented with fibroblasts can reproduce the heterogeneity to prevent changes in regulation [91]. Another approach is xenografting tumors derived from single cells in permissive mice as was described by Quintana *et al*. However, expanding the amount of tumor material was shown to be influenced by the host system since tumors that grew out were heterogeneous in gene expression [92].

Single cell sequencing directly from tumor material does not change the phenotype of cells [93]. Advances have been made towards single cell sequencing of both DNA and RNA and single cell proteomics. However, these techniques cannot be performed in the same cell [90, 94]. Sequencing single cells throughout entire tumors has shown that heterogeneity of tumors can be detected and clonal expansion can be predicted [95]. Nevertheless, tumors also contain other cell types that obstruct the analysis, such as blood vessels, stromal tissue, necrotic tissue and cells involved in the immune response. The presence of different cell types could lead to the detection of pathways not involved in resistance but rather other responses such as immune responses and anti-apoptotic pathways. This effect is stimulated by vemurafenib, which has been shown to induce the immune response [96].

Quality of acquired data

The model described in this review assumes that the data obtained at different levels and between different samples is equal. Sensitivity of data acquisition influences the quality of the data and should be similar between datasets. However, in practice this cannot be assured and differences in quality of data acquisition will cause discrepancies in integration [97].

Obtaining enough data to generate an integrative model from a single cell is currently impossible [90, 94]. Material derived from single cells needs to be amplified in order to perform sequencing, creating a bias in the abundance of fragments [93]. Additionally, single cell transcriptomics results in a noisy expression signature decreasing the quality of the data [98]. Therefore, it is not possible to generate perfectly accurate data from single cells at this moment.

Integration of data needs to be performed over several cells since currently the quantity and quality of single cell data is not sufficient. However, acquiring data over regions in the tumor could cause differences in abundance between cells but can also change the correlations between layers of data. By analyzing the proteome of several human tissues it was found that protein abundance varies between cells of the same type [99]. Additionally, uncommon alterations that might cause drug resistance can be missed since the acquired data will be averaged over a colony of cells [100].

Feasibility

The size of the genomic, transcriptomic and proteomic data requires extensive analysis and expensive data acquisition raising the question whether modelling is feasible. For instance, sequencing an entire human genome currently costs more than $ 1000 [101]. To study resistance genomic, transcriptomic and proteomic data needs to be generated before and after resistance occurs. As a model system melanoma is not the most suitable since resistance occurs after six months introducing more variation. In colon cancer patients drug resistance occurs almost instantly [9]. Because of the quick response colon cancer samples can be used instead of melanoma samples to accelerate the detection of resistance mechanisms, making the experiment more reproducible. Using single biopsies for whole-genome sequencing from, for instance, 20 colon cancer patients would require approximately $ 40.000. Adding transcriptomics, proteomics and phosphoproteomics data would be even more expensive but remains feasible. However, integrating all 17 modules will not be feasible. Additionally, detecting new mechanisms will require more than 20 patients to be enrolled since more than 20 mechanisms of drug resistance have currently been identified [3, 4].

Using single biopsies before and after treatment does not provide all the information to correct for heterogeneity of the tumor. The lack of quantity and quality of data used for integration indicates that creation of a model requires several biopsies to be taken from the tumor. Using this approach, data can be generated for multiple regions of a tumor and analyzed separately to acquire information about different mechanisms [16, 90]. However, protein interactions vary over time requiring a time series of biopsies as well. The amount of material obtained from several small biopsies is currently not sufficient to obtain genomics, transcriptomics and proteomics data [90, 94]. To expand the amount of material it is possible to culture tumor cells from patients in an organoid culture or xenografting cells into permissive mice [91, 92]. Hypothetically, if a serial biopsy is taken from 20 patients at 10 different timepoints from 10 regions in the tumor this will result in 2000 biopsies and cultures. In total this would require more than 2 million dollar for merely the genomic information and is therefore not feasible to perform for all processes described above.

Clinical applications

Integration of different layers of data in a tumor is an essential step in understanding cancer drug resistance. Novel mechanisms discovered through integrative approaches will lead to novel insights and treatment combinations. Alternatively, if applied on a greater scale it can be used to validate mechanisms that occur in patients since not all mechanisms were found to confer resistance in patients. For instance, it was proposed that secondary mutations in BRAF can confer resistance yet this mechanism has never been observed in patients [4]. Developing drugs against mechanisms that do not occur or occur rarely is costly and only provides treatment benefit for a small population of patients. An important addition to the model would be to include phosphoproteomics data to find which pathway is activated and in which pathways relief of feedback inhibition take place. This is of great clinical interest since temporary withdrawal of the drug (drug holiday) has been shown to make tumors regress and is most likely caused by reintroducing feedback inhibition [102].

The use of the model will still be limited for diagnostics and treatment as it is uncertain how tumors develop spatially and over time. Acquiring the data necessary to investigate this is currently not feasible. It is possible to obtain data on processes occurring at a position in a tumor at a specific time. This data could be used to choose a treatment strategy that is effective for that position at that time. By integrating genomic, transcriptomic and proteomic data from individual subclones in the tumor or by obtaining omics data over time predictions about the cause of clonal expansion can be made. Yet, a significant limitation to this approach remains the lack of tumor material.

Improving treatment strategies for individual patients, as was shown in HIV infection, can be achieved by using biomarkers. For HIV infection it is sufficient to sequence the virus in order to detect which drugs are effective [87]. For personalized treatment biomarkers can be used to find occurring resistance mechanisms in patients instead of integrating all layers of complexity. Using DNA as a biomarker is most likely not as effective as using expression and/or phosphoproteomics data. The integrative approach described here could detect which changes best predict treatment response by applying it to multiple patients.

An integrative approach can provide more information that can be used for diagnostics and treatment. In theory, it is possible to integrate multiple layers of omics into a single model. However, technical advances need to be made to make data acquisition feasible. An integrative approach will not be enough to prevent the progression of tumors altogether, but it is likely to be invaluable in devising treatment strategies for individual patients in the future.

Acknowledgements

I hereby thank Edwin Cuppen for allowing me to write this thesis and also for being the second reviewer on the report. In addition, I would like to thank Joep de Ligt for the opportunity to write under his supervision. The suggestions and support have helped me develop my writing skills further and increased my enthusiasm for systems biology.

References

1. Larkin, J., et al., *Vemurafenib in patients with BRAF(V600) mutated metastatic melanoma: an open-label, multicentre, safety study.* Lancet Oncol, 2014. **15**(4): p. 436-44.
2. Sun, C., et al., *Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma.* Nature, 2014. **508**(7494): p. 118-22.
3. Aplin, A.E., F.M. Kaplan, and Y. Shao, *Mechanisms of resistance to RAF inhibitors in melanoma.* J Invest Dermatol, 2011. **131**(9): p. 1817-20.
4. Sullivan, R.J. and K. Flaherty, *MAP kinase signaling and inhibition in melanoma.* Oncogene, 2013. **32**(19): p. 2373-9.
5. Karr, J.R., et al., *A whole-cell computational model predicts phenotype from genotype.* Cell, 2012. **150**(2): p. 389-401.
6. Low, T.Y., et al., *Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis.* Cell Rep, 2013. **5**(5): p. 1469-78.
7. Davies, H., et al., *Mutations of the BRAF gene in human cancer.* Nature, 2002. **417**(6892): p. 949-54.
8. Wan, P.T., et al., *Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF.* Cell, 2004. **116**(6): p. 855-67.
9. Prahallad, A., et al., *Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR.* Nature, 2012. **483**(7387): p. 100-3.
10. Chapman, P.B., et al., *Improved survival with vemurafenib in melanoma with BRAF V600E mutation.* N Engl J Med, 2011. **364**(26): p. 2507-16.
11. Kim, K.B., et al., *Phase II study of the MEK1/MEK2 inhibitor Trametinib in patients with metastatic BRAF-mutant cutaneous melanoma previously treated with or without a BRAF inhibitor.* J Clin Oncol, 2013. **31**(4): p. 482-9.
12. Wagle, N., et al., *Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling.* J Clin Oncol, 2011. **29**(22): p. 3085-96.
13. Nazarian, R., et al., *Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation.* Nature, 2010. **468**(7326): p. 973-7.
14. Vergani, E., et al., *Identification of MET and SRC activation in melanoma cell lines showing primary resistance to PLX4032.* Neoplasia, 2011. **13**(12): p. 1132-42.
15. Straussman, R., et al., *Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion.* Nature, 2012. **487**(7408): p. 500-4.
16. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
17. Kaufmann, W.K., et al., *Mechanisms of chromosomal instability in melanoma.* Environ Mol Mutagen, 2014.
18. Poulikakos, P.I., et al., *RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E).* Nature, 2011. **480**(7377): p. 387-90.
19. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
20. Balbin, O.A., et al., *Reconstructing targetable pathways in lung cancer by integrating diverse omics data.* Nat Commun, 2013. **4**: p. 2617.
21. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
22. Henikoff, S., *Beyond the central dogma.* Bioinformatics, 2002. **18**(2): p. 223-5.
23. Mattick, J.S., *Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms.* Bioessays, 2003. **25**(10): p. 930-9.
24. Ryan, C.J., et al., *High-resolution network biology: connecting sequence with function.* Nat Rev Genet, 2013. **14**(12): p. 865-79.
25. Mullins, J.G., *Structural modelling pipelines in next generation sequencing projects.* Adv Protein Chem Struct Biol, 2012. **89**: p. 117-67.

26. Ning, K., D. Fermin, and A.I. Nesvizhskii, *Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data.* J Proteome Res, 2012. **11**(4): p. 2261-71.
27. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
28. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
29. Tang, W., Y. Fei, and M. Page, *Biological significance of RNA editing in cells.* Mol Biotechnol, 2012. **52**(1): p. 91-100.
30. Omenn, G.S., R. Menon, and Y. Zhang, *Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology.* J Proteomics, 2013. **90**: p. 28-37.
31. Faghihi, M.A. and C. Wahlestedt, *Regulatory roles of natural antisense transcripts.* Nat Rev Mol Cell Biol, 2009. **10**(9): p. 637-43.
32. Kiefer, F., et al., *The SWISS-MODEL Repository and associated resources.* Nucleic Acids Res, 2009. **37**(Database issue): p. D387-92.
33. Pieper, U., et al., *ModBase, a database of annotated comparative protein structure models and associated resources.* Nucleic Acids Res, 2014. **42**(Database issue): p. D336-46.
34. Prasad, N.K., et al., *Structural and docking studies of Leucaena leucocephala Cinnamoyl CoA reductase.* J Mol Model, 2011. **17**(3): p. 533-41.
35. Gajadhar, A.S. and F.M. White, *System level dynamics of post-translational modifications.* Curr Opin Biotechnol, 2014. **28C**: p. 83-87.
36. Myllykangas, S., et al., *Classification of human cancers based on DNA copy number amplification modeling.* BMC Med Genomics, 2008. **1**: p. 15.
37. Dostie, J. and J. Dekker, *Mapping networks of physical interactions between genomic elements using 5C technology.* Nat Protoc, 2007. **2**(4): p. 988-1002.
38. Hesselberth, J.R., et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods, 2009. **6**(4): p. 283-9.
39. Smith, Z.D., et al., *High-throughput bisulfite sequencing in mammalian genomes.* Methods, 2009. **48**(3): p. 226-32.
40. Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature, 2005. **435**(7043): p. 834-8.
41. Kervestin, S. and A. Jacobson, *NMD: a multifaceted response to premature translational termination.* Nat Rev Mol Cell Biol, 2012. **13**(11): p. 700-12.
42. Bordbar, A., et al., *Constraint-based models predict metabolic and associated cellular functions.* Nat Rev Genet, 2014. **15**(2): p. 107-20.
43. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.
44. Nakakuki, T., et al., *Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics.* Cell, 2010. **141**(5): p. 884-96.
45. Zhang, J., et al., *Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data.* BMC Syst Biol, 2013. **7 Suppl 2**: p. S4.
46. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.
47. Shen, R., N.C. Goonesekere, and C. Guda, *Mining functional subgraphs from cancer protein-protein interaction networks.* BMC Syst Biol, 2012. **6 Suppl 3**: p. S2.
48. Brun, C., et al., *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.* Genome Biol, 2003. **5**(1): p. R6.
49. Zhang, S., et al., *Determining modular organization of protein interaction networks by maximizing modularity density.* BMC Syst Biol, 2010. **4 Suppl 2**: p. S10.

50. Vandin, F., E. Upfal, and B.J. Raphael, *Algorithms for detecting significantly mutated pathways in cancer.* J Comput Biol, 2011. **18**(3): p. 507-22.
51. Wilkinson, D.J., *Bayesian methods in bioinformatics and computational systems biology.* Brief Bioinform, 2007. **8**(2): p. 109-16.
52. Sachs, K., et al., *Bayesian network approach to cell signaling pathway modeling.* Sci STKE, 2002. **2002**(148): p. pe38.
53. Jia, P. and Z. Zhao, *VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data.* PLoS Comput Biol, 2014. **10**(2): p. e1003460.
54. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes.* Hum Mol Genet, 2010. **19**(R2): p. R131-6.
55. Snyder, M., J. Du, and M. Gerstein, *Personal genome sequencing: current approaches and challenges.* Genes Dev, 2010. **24**(5): p. 423-31.
56. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.
57. Mamanova, L., et al., *FRT-seq: amplification-free, strand-specific transcriptome sequencing.* Nat Methods, 2010. **7**(2): p. 130-2.
58. Anders, S., A. Reyes, and W. Huber, *Detecting differential usage of exons from RNA-seq data.* Genome Res, 2012. **22**(10): p. 2008-17.
59. Dehouck, Y., et al., *PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality.* BMC Bioinformatics, 2011. **12**: p. 151.
60. Ward, L.D. and M. Kellis, *Interpreting noncoding genetic variation in complex traits and human disease.* Nat Biotechnol, 2012. **30**(11): p. 1095-106.
61. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-43.
62. Pollack, J.R., et al., *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12963-8.
63. Yang, J., et al., *Detection of candidate tumor driver genes using a fully integrated Bayesian approach.* Stat Med, 2014. **33**(10): p. 1784-800.
64. Geiger, T., J. Cox, and M. Mann, *Proteomic changes resulting from gene copy number variations in cancer cells.* PLoS Genet, 2010. **6**(9): p. e1001090.
65. Havugimana, P.C., et al., *A census of human soluble protein complexes.* Cell, 2012. **150**(5): p. 1068-81.
66. Lau, C., et al., *ERBB4 mutation analysis: emerging molecular target for melanoma treatment.* Methods Mol Biol, 2014. **1102**: p. 461-80.
67. Cawley, G.C. and N.L. Talbot, *Kernel learning at the first level of inference.* Neural Netw, 2014. **53**: p. 69-80.
68. Larjo, A., I. Shmulevich, and H. Lahdesmaki, *Structure learning for Bayesian networks as models of biological networks.* Methods Mol Biol, 2013. **939**: p. 35-45.
69. Liu, Z., B. Malone, and C. Yuan, *Empirical evaluation of scoring functions for Bayesian network model selection.* BMC Bioinformatics, 2012. **13 Suppl 15**: p. S14.
70. Berger, E., et al., *HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data.* PLoS Comput Biol, 2014. **10**(3): p. e1003502.
71. Wagner, J.R., et al., *The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts.* Genome Biol, 2014. **15**(2): p. R37.
72. Adler, A.S., et al., *An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth.* Genes Dev, 2014. **28**(10): p. 1068-84.
73. Marquardt, S., D.Z. Hazelbaker, and S. Buratowski, *Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species.* Transcription, 2011. **2**(3): p. 145-154.
74. Kim, D., et al., *Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction.* Methods, 2014.

75. Jiang, X., et al., *A bayesian method for evaluating and discovering disease loci associations.* PLoS One, 2011. **6**(8): p. e22075.
76. Spirtes, P., C.N. Glymour, and R. Scheines, *Causation, prediction, and search*. Vol. 81. 2000: MIT press.
77. Yen, E.A., et al., *Exploration of the dynamic properties of protein complexes predicted from spatially constrained protein-protein interaction networks.* PLoS Comput Biol, 2014. **10**(5): p. e1003654.
78. Whittaker, S., et al., *Gatekeeper mutations mediate resistance to BRAF-targeted therapies.* Sci Transl Med, 2010. **2**(35): p. 35ra41.
79. Tsai, J., et al., *Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity.* Proc Natl Acad Sci U S A, 2008. **105**(8): p. 3041-6.
80. Lito, P., et al., *Relief of profound feedback inhibition of mitogenic signaling by RAF inhibitors attenuates their activity in BRAFV600E melanomas.* Cancer Cell, 2012. **22**(5): p. 668-82.
81. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
82. Goncalves, E., et al., *Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models.* Mol Biosyst, 2013. **9**(7): p. 1576-83.
83. Antoniewicz, M.R., *Dynamic metabolic flux analysis--tools for probing transient states of metabolic networks.* Curr Opin Biotechnol, 2013. **24**(6): p. 973-8.
84. Altelaar, A.F., J. Munoz, and A.J. Heck, *Next-generation proteomics: towards an integrative view of proteome dynamics.* Nat Rev Genet, 2013. **14**(1): p. 35-48.
85. Lerman, J.A., et al., *In silico method for modelling metabolism and gene product expression at genome scale.* Nat Commun, 2012. **3**: p. 929.
86. Zheng, S., M.G. Chheda, and R.G. Verhaak, *Studying a complex tumor: potential and pitfalls.* Cancer J, 2012. **18**(1): p. 107-14.
87. Van der Borght, K., et al., *Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations.* BMC Bioinformatics, 2011. **12**: p. 386.
88. van Westen, G.J., et al., *Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data.* PLoS Comput Biol, 2013. **9**(2): p. e1002899.
89. Pisco, A.O., et al., *Non-Darwinian dynamics in therapy-induced cancer drug resistance.* Nat Commun, 2013. **4**: p. 2467.
90. Shi, Q., et al., *Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells.* Proc Natl Acad Sci U S A, 2012. **109**(2): p. 419-24.
91. De Witt Hamer, P.C., et al., *The genomic profile of human malignant glioma is altered early in primary cell culture and preserved in spheroids.* Oncogene, 2008. **27**(14): p. 2091-6.
92. Quintana, E., et al., *Efficient tumour formation by single human melanoma cells.* Nature, 2008. **456**(7222): p. 593-8.
93. Junker, J.P. and A. van Oudenaarden, *Every cell is special: genome-wide studies add a new dimension to single-cell biology.* Cell, 2014. **157**(1): p. 8-11.
94. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis.* Nat Methods, 2014.
95. Francis, J.M., et al., *EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing.* Cancer Discov, 2014.
96. Koya, R.C., et al., *BRAF inhibitor vemurafenib improves the antitumor activity of adoptive cell immunotherapy.* Cancer Res, 2012. **72**(16): p. 3928-37.
97. Tarca, A.L., G. Bhatti, and R. Romero, *A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity.* PLoS One, 2013. **8**(11): p. e79217.
98. Grun, D., L. Kester, and A. van Oudenaarden, *Validation of noise models for single-cell transcriptomics.* Nat Methods, 2014. **11**(6): p. 637-40.

99. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.

100. Campbell, P.J., et al., *Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing.* Proc Natl Acad Sci U S A, 2008. **105**(35): p. 13081-6.

101. Morey, M., et al., *A glimpse into past, present, and future DNA sequencing.* Mol Genet Metab, 2013. **110**(1-2): p. 3-24.

102. Das Thakur, M., et al., *Modelling vemurafenib resistance in melanoma reveals a strategy to forestall drug resistance.* Nature, 2013. **494**(7436): p. 251-5.