biomolecules

MDPI

Review

The Road Not Taken with Pyrrole-Imidazole Polyamides: Off-Target Effects and Genomic Binding

Jason Lin * and Hiroki Nagase *

Laboratory of Cancer Genetics, Chiba Cancer Center Research Institute, Chiba 260-8717, Japan
* Correspondence: linjason@chiba-cc.jp (J.L.); hnagase@chiba-cc.jp (H.N.)

Received: 21 February 2020; Accepted: 19 March 2020; Published: 3 April 2020

Abstract: The high sequence specificity of minor groove-binding *N*-methylpyrrole-*N*-methylimidazole polyamides have made significant advances in cancer and disease biology, yet there have been few comprehensive reports on their off-target effects, most likely as a consequence of the lack of available tools in evaluating genomic binding, an essential aspect that has gone seriously underexplored. Compared to other *N*-heterocycles, the off-target effects of these polyamides and their specificity for the DNA minor groove and primary base pair recognition require the development of new analytical methods, which are missing in the field today. This review aims to highlight the current progress in deciphering the off-target effects of these *N*-heterocyclic molecules and suggests new ways that next-generating sequencing can be used in addressing off-target effects.

Keywords: *N*-heterocycles; minor-groove binding; pyrrole-imidazole polyamides; genomics; chemical biology

1. Introduction

N-methylpyrrole-*N*-methylimidazole polyamides (pyrrole-imidazole polyamides, PIPs) possess the incredible ability to interact and bind the DNA minor groove with programmable specificities exceeding distamycin (Figure 1a), the molecule that sparks the inspiration behind this field. These heterocycles bind and are able to differentiate their respective nucleosides to single hydrogen-bond precisions (Figure 1b, [1]) and are typically generated by solid-phase [2] or, at times, solution-phase synthesis [3]; the production of PIPs can be readily scaled up to milligram scales for a variety of different biological applications. Recent studies on PIP have been investing on the possible applications in silencing "undruggable" biological targets that are difficult to inhibit at the protein level. Those so-called undruggable targets present challenges of drug sensitivity and resistance because of their intracellular localization and numerous associations with other proteins and co-factors in macromolecular complexes [4]. In contrast, upon minor-groove binding, these polyamides will effectively silence their respective marks by disrupting transcription, bypassing issues associated with the usual approach of inhibition via the binding pocket or surface access. This unique biochemical feature makes PIPs especially attractive as a medium of mitigating the virulence of oncological targets, a number of which have arisen as a consequence of tens to millions of genetic and chromosomal aberrations harbored within the cancer genome [5,6].

Currently, a significant portion of research in PIPs is centralized on lead developments; this relocates the focus squarely onto the exploration of possible targets in diseases and biological systems. There is significant focus on the chemistry and the effects of chemical modifications on these polyketides in terms of improving their pharmacokinetic parameters, such as the tuning of lipophilicity and motif recognition. While most of these results have been promising, an area severely underinvestigated is the possibility of off-target bindings and their potential implications. These off-target-binding events lead to the rise of adverse effects as other genes are transcriptionally impacted, subsequently triggering changes to other biological processes which may be totally unrelated to the target of interest [7]. It is but a certain mathematical possibility that all PIPs currently in development will bind multiple targets as a consequence of their mode of motif recognition. Suppose the probability that a particular *k*-mer occupies a particular genomic location is $(4k)^{-1}$; accounting for DNA being double-stranded, we can reasonably posit that the expected number of *k*-mer repeats in the human genome is approximately $2N^2/4^k$, with *N* being roughly three billion. For the expectation to fall close to 0, a back-of-the-envelope calculation would reveal that, at a minimum, we need to achieve a base recognition rate beyond 32 to 36 bases. With siRNAs, which tend to be 20–25-bases-long, the expectation that there is only one binding site is a mathematical impossibility; PIPs and their usual span of recognition being 5 to 10 nucleotides are also out of the range of uniqueness. While efforts to lengthen PIP recognition by tandem conjugation have succeeded [8,9], with the molecular size, it very quickly reaches near the ceiling of the nuclear diffusion limit of ~10 nm for macromolecules [10,11].

This "unfortunate" aspect of being unable to reach uniqueness, however, does not discount the usefulness of PIPs. Even with CRISPR/Cas9 systems, there is nonetheless a need to extend guide RNAs beyond the same limit of uniqueness, not to mention the myriad of undesired effects such as exon-skipping and truncated expressions [12] or recent controversies on using such a genome-editing tool on human subjects. These concerns squarely place PIPs back into the focus as therapeutic alternatives for various diseases, but along the same vein, one cannot simply disregard the likelihood of off-target binding and the resulting consequence of these binding events. Unlike other nitrogen heterocycles that find their way onto the list of 59% of approved drugs in the United States [13] or DNA-binding drugs such as cisplatin, doxorubicin or cryptolepine [14], PIPs inherently "bundle" programmability, with the benefit of having fewer off-target sites by the virtue of their lengthened motif recognition; yet, those other molecules see extensive research in their adverse indications, while the field for PIPs is paradoxically barren. Why are there so few willing to venture down this path? This review aims to discuss the feasibility of PIPs as pharmaceutical leads, outline some of the current challenges in the evaluation of off-target binding in these molecules and comment on the possible future directions in exploring off-target binding.

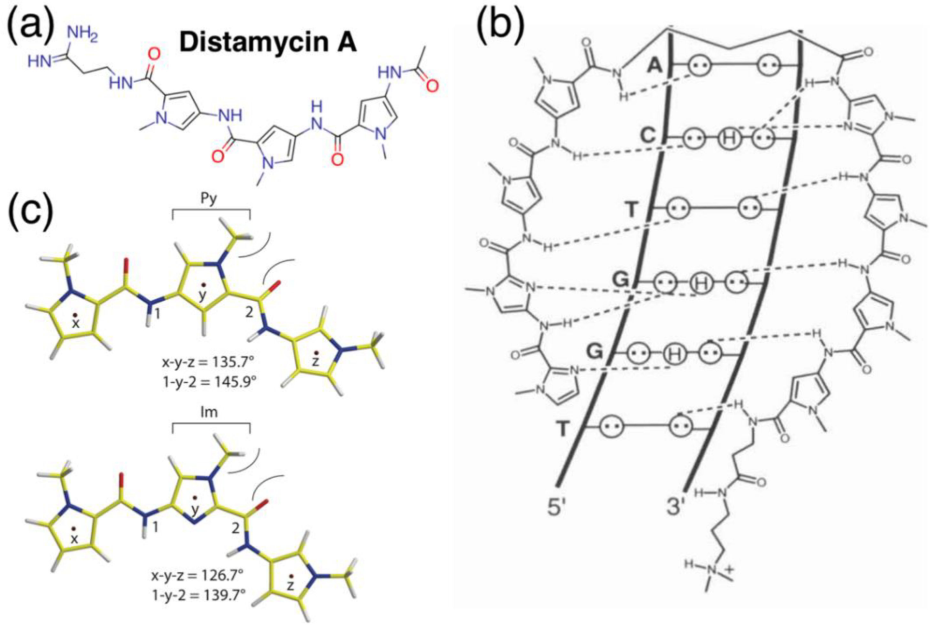

Figure 1. *Cont.*

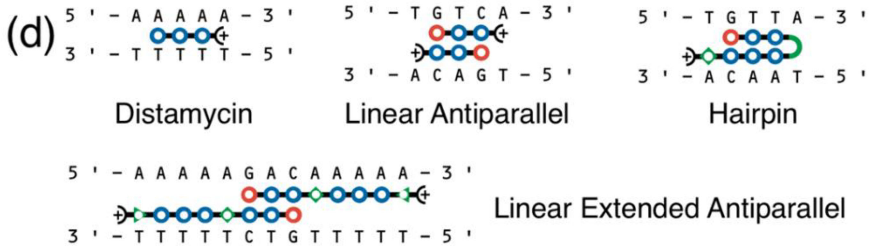

Figure 1. Schematics of pyrrole-imidazole polyamides and DNA binding. (**a**) Structure of distamycin A. (**b**) Illustrative model of hydrogen-bonding interactions between pyrrole-imidazole polyamides (PIPs) and the DNA minor groove; circles with dots represent lone pairs of purines and pyrimidines; circles containing an H represent the additional hydrogen on G; putative hydrogen bonds are indicated with dotted lines. Reprinted with permission from [1], © 2001 American Chemical Society. (**c**) Local conformation modeling of *N*-methylpyrroles (Py, top) and *N*-methylimidazoles (Im, bottom) in PIP backbones. Bond angles x-y-z and 1-y-2 are indicated directly below. Reprinted with permission from [15], © 2011 American Chemical Society. (**d**) Example of various binding models' PIP-DNA minor-groove complexes: linear antiparallel, hairpin and linear extended antiparallel. Red and blue circles represent imidazole and pyrrole rings, respectively; green triangles, diamonds, curved green lines and +) represent glycine, β-alanine, γ-aminobutyric acid and the *N*,*N*-dimethylaminopropylamide (Dp) capping, respectively. Reprinted from [16], © 1996, with permission from Elsevier.

2. Biological Applications of PIPs

Pyrrole-imidazole polyamides (PIPs) themselves occupy a corner of naturally inspired antibiotic mimetics; these *N*-heterocycles consist of a backbone of *N*-methylpyrrole and *N*-methylimidazole subunits (Figure 1c) assembled together by amide bonds [15] and are able to interact with the DNA minor groove in various structural conformations (Figure 1d), with the hairpin [16] configuration being one of the most commonly adopted. Despite the simplicity behind its chemistry, the amide bonds in PIP backbones permit the conjugation of a variety of different reactive moieties that diversify the structure and functionalities of PIPs. Their innate ability to bind the DNA minor groove allows them to be used to repress gene transcription via triggering RNA polymerase II inhibition [17], without having to resort to unstable or complex experimental procedures such RNAi or CRISPR/Cas. As such, by binding the minor groove, one can avoid the messiness of conventional small molecular inhibitors that work on protein surfaces. PIPs have been used in a variety of biological applications; in 2005, Beerman and colleagues successfully demonstrated the ability to cause DNA damage by conjugating 1-chloromethyl-5-hydroxy-1,2-dihydro-3H-benz(e)indole (CBI), an alkylating agent, to disrupt and inhibit SV40 DNA replication [18]; conjugation of other functional groups have also led to the rise of additional functionalities for PIPs, such as artificial nucleases [19], gene switches [20], transcriptional regulators via histone acetyltransferase activators [21] and imaging probes [22] upon the conjugation of fluorescent moieties (Figure 2a) or radioisotopes [23]. Even without those modifications, PIPs are still able to retain their high specificity, unlike mainstream DNA-damaging agents such as cisplatin [17].

The relative ease and robustness of synthesis, as well as the capability to conjugate functional groups by the same amide chemistry, is perhaps the most important feature that has brought PIPs to the forefront of chemical biology. The ability to target select regions of the genome and exert specific actions such as transcriptional disruption, epigenetic reprogramming or the recruitment of transcriptional elements has led to the development of a number of unique and interesting biological applications for PIPs, especially in the areas of cancer and disease biology. This also theoretically simplifies and increases the throughput of the design process. PIPs have their ways in the inhibition of oncological targets such as MMP [24], constitutively active *KRAS* [25], a frequent driver gene in cancers such as colorectal, lung and pancreatic cancers, as well as members of the *RUNX* transcription factor family [26]. Notably, the PIP-targeting G12D/V mutation in *KRAS* effectively restrained tumor growth in a mouse xenograft model (Figure 2b). PIPs have also shown effectiveness in targeting genetic aberrations other than single-base mutations, such as the case with copy-amplified *MYCN* [27], an aberration in neuroblastoma, where appreciable DNA damage and induced apoptosis were observable in *MYCN*-amplified neuroblastoma mouse models. There have also been reports of PIPs suppressing transcription factors central to hypoxic gene expressions relevant in neovascularization and cancer metastasis [28], as well as demonstrating antiproliferative activity in prostate cancer cells via the modulation of the androgen receptor [29,30] in ways different from other drugs such as camptothecin, doxorubicin or etoposide [31]. Some of the other examples include the ability to interfere with TNF-α-inducible transcriptome [32] and the triggering of inflammatory necrotic cell death in cancer cells via calreticulin, ATP and HMGB1 signaling [33]. Beyond cell-level and mouse model experiments, preclinical trials in marmosets also show promising success of a PIP targeting the human TGF-β1 promoter in reducing scarring [34].

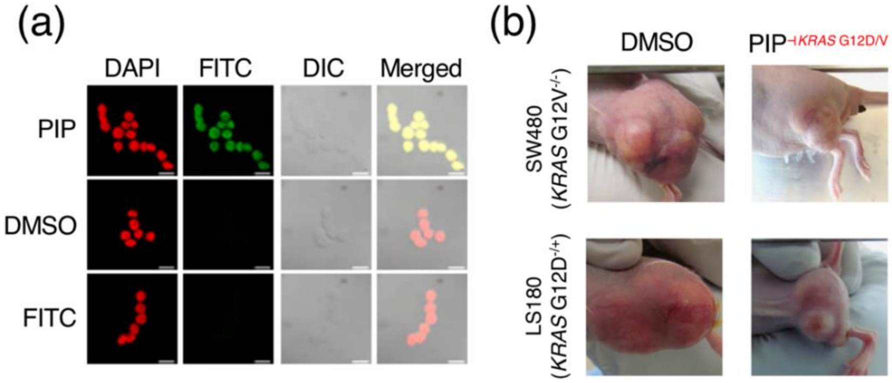

Figure 2. Biological applications of PIPs. (**a**) Example of in-situ cell imaging using a fluorescein isothiocyanate (FITC)-conjugated PIP; reproduced from [22], © 2018 with permission from Elsevier. (**b**) The coupling of 1-chloromethyl-5-hydroxy-1,2-dihydro-3H-benz(e)indole (CBI), an alkylating agent, to a PIP-targeting *KRAS* G12D/V mutation is able to reduce the size of tumors in human colorectal cancer LS180/SW480 xenograft mouse models; images shown here are representative of multiple specimens, with DMSO (left) as a control; reused from [25] by the author, © 2015 Springer Nature Limited.

3. Off-Target Effects and "The Road Not Taken"

Despite the stellar performance of PIPs in biological applications, this class of molecules has yet to see commercial success as leads in drug discovery. In theory, one can systematically design, synthesize and test a larger number of PIPs with relative ease, possibly even over compounds derived from combinatorial chemistry. With their specificity and unique ability to penetrate nucleus and target DNA compared to conventional broad-spectrum chemotherapy agents that are cytotoxic compounds with a large number of adverse indications, it is not unrealistic to expect the presence of multiple PIPs in clinical trials or, perhaps, even on the market today. There, however, seems to be little movement for PIPs in pharmaceutical chemistry. Furthermore, there have been few reports dedicated to the discussion of off-target effects, one of the most important factors in drug discovery; nearly 50% of the attritions during the United States Food and Drug Administration Phase II clinical trials are due to the occurrence of toxic side effects [35]. Understandably, most would prefer to steer conversations away from topics that could quickly turn to discount their own work, leading to tremendous undergrowth down this seldom-traversed route. Unfortunately, this remains a critical area of discussion, especially during lead discovery and optimization. Performing research in this area, however, is neither exactly pleasant nor encouraging. What, then, are off-target effects, and why do researchers not elect to investigate this path less traveled?

3.1. Defining and Evaluating "Off-Target" Effects

In the most general sense, binding to any genomic location other than a polyamide's intended site can be considered off-target, and in the human genome, it is a mathematical certainty that PIPs will have off-target effects. This is not a problem strictly related to PIPs, however; even revolutionary genome-editing systems like CRISPR/Cas are affected by this phenomenon; a recently discovered artifact in CRISPR/Cas, for even, is that genomic deletions induced by CRISPR may lead to unintended exon skipping and the production of aberrant proteins [12]. To attenuate these undesirable effects requires that we first understand *what* those effects are; unfortunately, our current knowledge in the subject matter is practically nonexistent. If we could understand these off targets, we could then design new PIP candidates that try not to target these genomic regions or make chemical modifications that alter the affinity of PIP-DNA ligand interactions. Similarly, by understanding these off targets, we may be able to reposition PIPs for other applications in a manner not unlike others have done in pharmaceutical chemistry, such as Zidovudine [36] or Viagra® [37].

Currently, evaluating off-target effects primarily involves the use of various biochemical and biophysical assays (see examples in Table 1) to compare PIP binding with full-match and mismatch nucleotides. By understanding the differences in recognition affinity, we can deduce to some extent the likelihood of a PIP's cross-reactivity with unintended regions of the genome, along with valuable biological insights on the mode of action of PIPs in vivo, such as the possible dependence on chromatin structure and histone modifications on DNA accessibility [38] to help characterize the properties of a candidate PIP. These assays typically provide us with some measures of quantitative differentiation, e.g., equilibrium association/dissociation rates (Figure 3a) via surface plasmon resonance (SPR) or preferential binding rates via digital polymerase chain reactions [39]. For instance, structural flexibilities near the termini of the polyamide backbone will substantially alter the association and dissociation rate constants to shift the binding affinity of the PIP towards mismatched DNA [40,41]. Qualitatively, gel-shift electrophoresis can also provide a preliminary and inexpensive assessment of the relative affinity of PIP-DNA ligands. Circular dichroism experiments (Figure 3b) can also elucidate additional biophysical insights, such as the influence on DNA binding as the minor groove becomes narrowed due to changes in ionic strength that ultimately restricts groove size and alters the electrostatic repulsion of the negatively charged phosphate groups [42]; in some cases, circular dichroism may also be used to determine whether particular functional groups have successfully been delivered or activated [43] to the proximity of target DNA.

Table 1. Common biophysical assays used to assess pyrrole-imidazole polyamide (PIP)-DNA binding in situations where one DNA sequence interacts to multiple PIPs or vice versa. "Response factor" describes the primary type of measurement collected for evaluating PIP-binding specificity in relation to off-target binding.

Assay	Purpose	Response Factor	Frequency of Use [1]
Melting temperature	Comparison of PIP-dsDNA complex stabilities	ΔT_m	+++
Gel-shift electrophoresis	Detection of whether a candidate PIP interacts with DNA containing a target-binding motif or otherwise	Changes in gel mobility	+++
Digital PCR	Comparison of the relative presence of PIP-DNA complexes by affinity enrichment, e.g., streptavidin-biotin interactions	Fold enrichment of PIP-DNA species	+
Surface plasmon resonance	Evaluation of adsorption and desorption rates to determine the kinetics of PIP-DNA and relative binding strength	Kinetic rate (e.g., k_a or k_d) and equilibrium constants (e.g., K_D)	++
Circular dichroism	Monitoring of changes in DNA conformation structure upon PIP binding to discern differences in strength of the interaction with the minor groove	Changes in spectrometric signals within certain wavelengths	+

[1] Relative frequency an assay is used to discern differences in PIP-DNA interactions; "+++" is the most common, followed by "++" and "+".

Piecing together these molecular level clues about how PIPs behave, we may be able to derive a crude understanding of how these biochemical details will influence off-target binding; indeed, undoubtedly, changes in the local PIP-DNA ligand conformation will weaken this interaction and perchance destabilize nearby transcription elements, subsequently disrupting the transcription, although we may not be able to decipher what sort of phenotypic changes these can lead to. Nonetheless, one can safely posit that epigenetic changes are intimately linked to the rise of off-target effects. This is well-corroborated with the fact that small conformational distortions are sufficient to retard RNA polymerase II [17] and potentially displace histones [44]. Evidence also suggests that lipophilicity of a PIP is associated in tumor-directed delivery in situ [22], albeit there may be some dependence on the type and origin of tumors [45]. We can thus potentially expand our synthetic repertoire to introduce novel modifications to the existing PIP design toolbox to incorporate these improvements, but there are still caveats to take note of, not to mention again that these assays may only have provided us with qualitative results. For instance, further increasing a PIP's lipophilicity may increase its propensity to aggregate, reducing its biocompatibility and restricting its ability for tissue penetration. Despite reports that the aggregation propensity of PIPs should not contribute to biological activity and may not be a critical concern in pharmacokinetic analyses [46] and conflicting reports suggesting possible liver toxicity with certain hairpin modifications [47] in mouse models at a dose of ~10 mg/kg, one needs to be aware that there are still limits to the amount of chemical modifications to apply to PIPs, and other resolutions need to be attempted to address this complex problem.

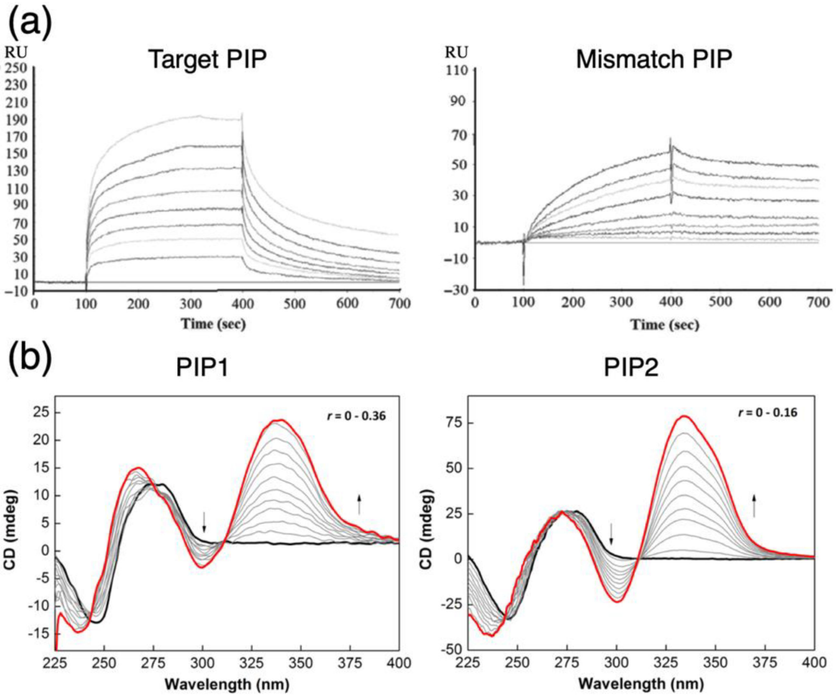

Figure 3. *Cont.*

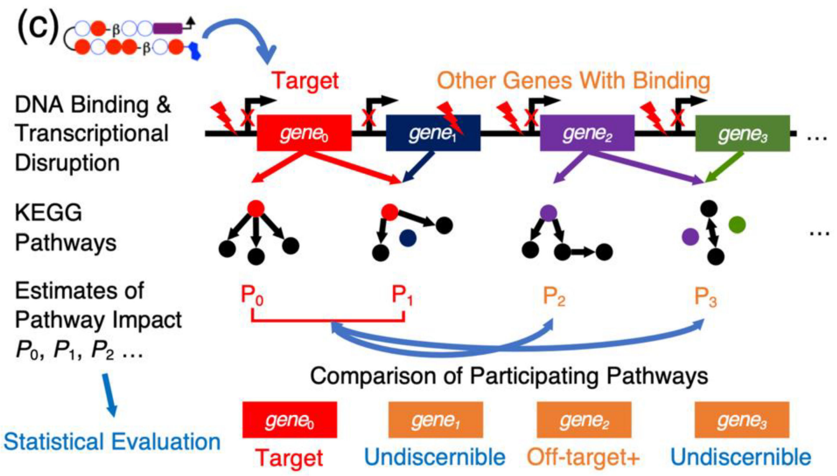

Figure 3. Analyzing off-target effects of PIPs with biophysical assays and expression profiling. (**a**) Illustrative use of surface plasmon resonance to compare relative binding specificities of PIPs to DNA containing a particular motif. Evaluation of the polyamides (“Target PIP” and “Mismatch PIP”) rely on differences in the association-dissociation kinetics; equilibrium constants can be estimated based on the trajectories in the sensorgram to elucidate further details about how a particular design variation will affect binding and the extent of binding to mismatched sequences; reproduced from [48], © 2015 (license terms CC BY-NC-ND). (**b**) Illustrative use of circular dichroism to determine the relative difference in PIP-DNA interactions via the monitoring of local conformational changes in the minor groove. The presence of features such as isoelliptic points in the DNA-absorbing wavelength region (230–300 nm) and induced signals near 355 nm indicate a favorable interaction. Comparison of spectra of different PIP candidates allow comparative assessments to be made. Adapted with permission from [42], © 2015 American Chemical Society. (**c**) A proposed method of determining the candidate of off-target genes using expression profiling and the binding site (modified from [49], © 2019 (license terms CC BY); the method utilizes pathway information to reduce the dimensionality of expression arrays to estimate off-target effects at the pathway level (illustrated in the example as $P_0, \ldots, P_3$). These metrics could then be used to compute overall changes that may be deconvoluted later on to identify potential off targets (“Off-target+” in orange) from genes that were somehow impacted as a consequence of other on-target and off-target genes in the pathway (“Undiscernible” in blue).

Aside from a PIP’s biophysical properties, arguably the most critical aspect in understanding what off-target effects are is the frequency and likelihood of genomic-binding events. Whereas, with protein-level inhibitors, one may be able to piece together structural information and interactions with other proteins to infer drug indications [50]. We can estimate a PIP’s binding sites by searching for its recognition motif in a sequence file, but to understand the frequency (and likelihood), more fine-grained approaches are required. A likely choice in this case is expression microarrays, for one can easily evaluate simultaneously changes in an entire genome with a single chip. However, changes in the expression profile may tell us what unexpected genes are up- or downregulated as a consequence of PIP administration, but they provide only a stochastic snapshot of the dynamic profile and cannot infer what actual mechanisms lead to the rise of these regulatory changes. Since chemical stimuli trigger environmental or oxidative stress homeostasis, those changes we observed may simply be a nonspecific response to having been perturbed by the mere event of PIP diffusion, complicating our understanding of the question we wish to address; how does one formulate a solution to such a question, then? A recent proposal is to reformulate the question to convolute the expression change of individual genes into units of pathways [49] as a mean of reducing the dimensions of analytical complexity (Figure 3c); as the problem is simplified, this clarifies the null hypothesis to evaluate and compare pathways that we expect to be modified compared to other pathways that are ordinarily unaffected. This analysis, however, still makes two underlying assumptions that remain to be verified: first, PIP-binding sites in regions that do not affect a gene’s expression probably have no effect; second, binding to those other sites is equally likely, such that all genes affect the outcomes equally. To evaluate whether those two assumptions are applicable, we will need the power of next-generation sequencing at our disposal.

3.2. Application of Next-Generation Sequencing in Evaluating PIP Off-Target Effects

Next-generation sequencing (NGS) methods provide a conduit to probe a PIP's genomic interactions and, subsequently, address the question of off-target binding. When coupled with expression-profiling results, sequencing analysis of PIP binding can provide insights into the interaction of these minor groove-binding heterocycles in the genomic space via binding-site information and nearby epigenetic changes. There are currently two major directions of using NGS to address these questions: Bind-n-Seq, which primarily aims to resolve motif recognition, and Chem-seq, which utilizes affinity enrichment to capture information on genomic-binding sites.

3.2.1. Bind-n-Seq: Inferences on Binding Motifs

Bind-n-Seq is, as its name implies, is a sequencing method that involves the hybridization of a molecular target with DNA to form ligands and the subsequent sequencing of said ligands [51]. This method utilizes a library of barcoded synthetic and typically randomized oligonucleotides to interact with the molecular target in question. Upon affinity capture and hybridization, unbound nucleotides are removed by washing, pooled and sequenced (Figure 4a). Binding preferences in the form of motif logos or position-weight matrices are calculated based on the relative abundance of bound DNA. Bind-n-Seq studies are high in throughput and have become prevalent in PIP studies, as the resultant logos provide quantitative data between the relative affinity of a full-match motif compared to its mismatch counterparts. These sorts of experiments do not require a large number of reagents and allow simultaneous cross-comparisons of multiple biotinylated PIPs due to the adaptation of oligonucleotide barcoding; Dervan et al., for instance, were able to perform comparisons of eight PIPs with a single run of Bind-n-Seq [52]. This technique can also infer additional binding details out of the purview of biophysical assays; for example, the relative preference for binding orientations (for instance, see Figure 4b for the same study of polyamides targeting the 5′-CGCG-3′ sequence as the means of inhibiting CpG methylation). Bind-n-Seq is also compatible with PIPs functionalized with alkylating agents such as CBI that form covalent linkages with DNA via the adenine in the 3′ terminus of its recognition motif [53]; this makes the method a more preferable choice at times over SPR, as CBI-conjugated polyamides will react irreversibly with the immobilized oligonucleotide targets and prevent dissociation from the sensor chip.

Variations of Bind-n-Seq include Bind-n-Seq-MR, which incorporates reference sequences in an attempt to differentiate the binding preferences of *N*-methylpyrroles to As and Ts [54]. Another proposal in characterizing genomic-binding sites of PI polyamides is the method known as cognate site identifier, or CSI [55]. CSI (Figure 4c) uses motif-finding algorithms to generate a positional map, from motif similarity scores, of all possible PIP-binding sites available in the sequence landscape defined by probe information on a microarray, then infers binding preference with a microarray data. CSI is arguably a variation of Bind-n-Seq and generates motif preferences, which one can indirect infer-binding information from reads gathered from a whole-genome sequencing experiment, with the caveat that the actual peak-calling process is still dependent on the choice of peak callers, and this motif information is merely used for annotation. While PIPs have experimentally demonstrated superior differentiating capabilities between full-match and mismatch motifs, an annotation-based approach such as Homer [56] is not sufficiently reliable for characterizing binding sites. With most PIPs being relatively short in base recognition, there may be upwards of a million hypothetical sites for typical design permutations, and one cannot fully exclude the possibility of coincidental motif overlapping, especially in polyamides designed to target oncogenes in cancer cell lines, in which the mutation profile of the specimen being sequenced can contain large variations from the reference genome. The presence of sequencing artifacts and the level of background noise may also reduce the accuracy of peak callers, subsequently affecting the final annotation results. Nevertheless, while Bind-n-Seq and its variations are highly capable in inferring the binding specificities of PIPs, their reliance on synthetic nucleotide libraries limits the scope of the coverage landscape; to overcome this issue, Bind-n-Seq needs to be complemented by chromatin or whole-genome sequencing methods that will allow a PIP's actual in vivo binding locations to be thoroughly probed.

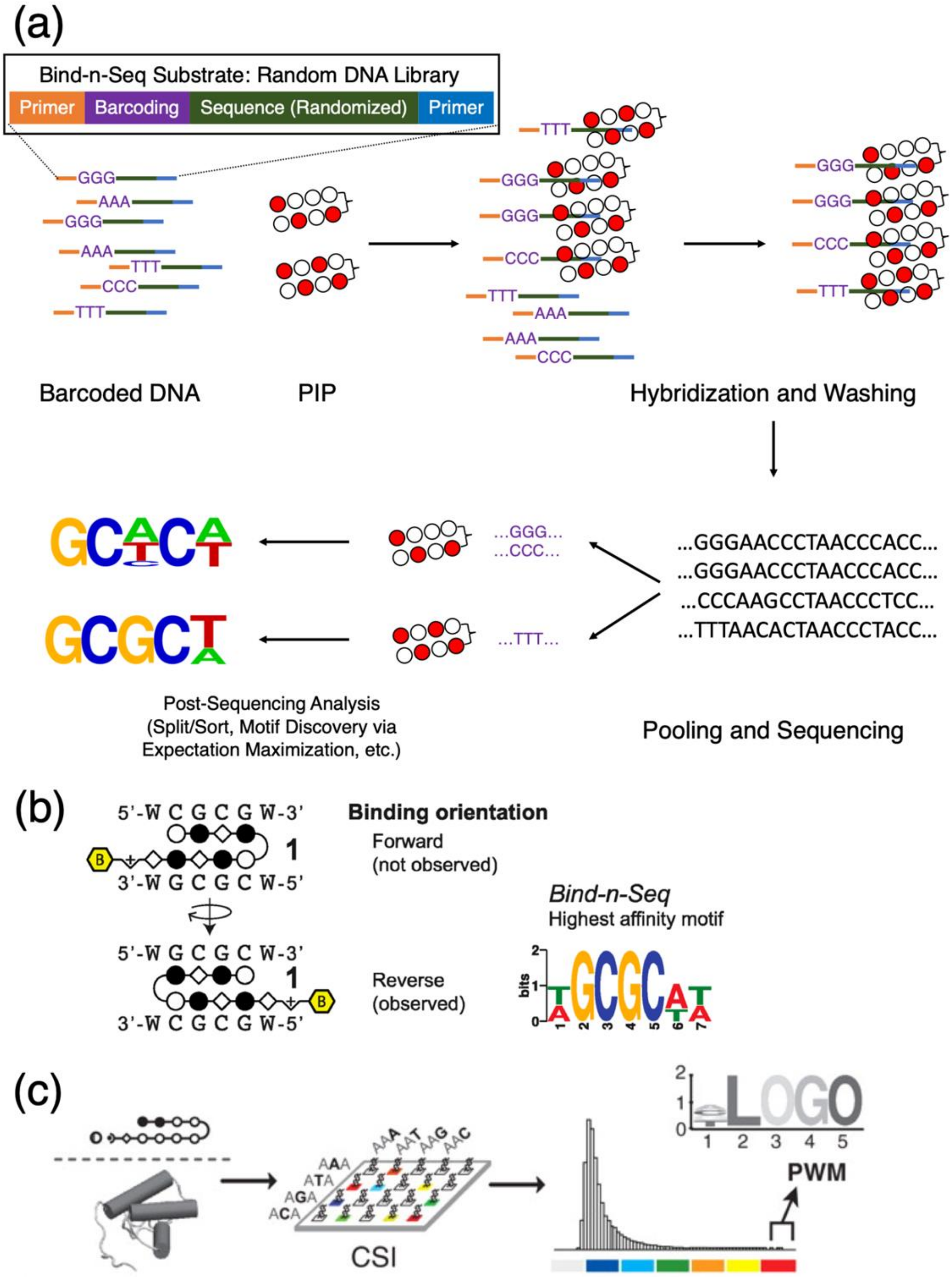

Figure 4. Mechanics of using Bind-n-Seq to elucidate the sequence specificity of PIPs. (**a**) Schematics of Bind-n-Seq, beginning with a random oligonucleotide library to create a possible sequence space for downstream analysis. (**b**) The ability to extract intricate details about PIP-binding modes with Bind-n-Seq; example shown here is the different in binding orientation. Reprinted with permission from [52], © 2014 American Chemical Society. (**c**) Cognate site identifier, or CSI, as a variation of a similar theme to identify binding motifs; motif enrichment results can then be transformed into position weight matrices (PWM) or presented as motif logo maps. Reproduced from [55], © 2010 National Academy of Sciences.

3.2.2. PIP Co-Sequencing Methods: Generating a Lemma for Possible Genomic Effects

PIPs indeed have been used in conjunction with typical ChIP-seq applications to probe changes in the epigenetic landscape, such as the effect of a targeted histone acetyltransferase activator at the *OCT-3/4* loci (Figure 5a, left) [21] or the ability for a polyamide to perturb the promoter region of *KLK3* [30] (Figure 5a, right), although these methods do not directly indicate the location of PIP binding. Rather, they require local sequencing reads to be either manually inspected or require multiple control experiments to be performed so a differential gain or loss of features can be observed. In the case of Han et al., up to four tracks were necessary to elucidate the effect of a histone acetyltransferase activator-conjugated PIP [21]. These companion ChIP-seq experiments provide implied evidence of PIPs interacting with genetic regulatory elements near said sites, and the results, while useful, are unable to corroborate mechanistic hypotheses such as the displacement of histone upon minor-groove binding. In contrast, the usefulness of these co-sequencing experiments is that they dispense a comprehensive image of the genome through changes in certain epigenetic features, e.g., H3K27Ac, in the presence of a PIP. With the loss or gain of local peaks as a proxy for the influence of a PIP, these types of experiments deliver indirect and limited information on a polyamide's exact genomic targets; to attain data on direct interacting targets necessitates the use of a PIP-centered sequencing method.

3.2.3. Chem-seq: Discovery of In Vivo Genomic Binding Sites

Chem-seq is the catch-all terminology of mapping genome-wide interactions and target sites of small molecules. While the earliest adoption of such a term was perhaps Anders et al. [57], the terminology has also been extended to PIPs, with several publications exclusively employing "Chem-seq" to refer to the sequencing of DNA fragments captured by PIP. This method primarily involves the use of alkylating PIPs functionalized with moieties such as a biotin enrichment tag, typically at one of the polyamide backbone termini or the hairpin turn, along with an indole-*seco*-CBI-reactive group (Figure 5b), to provide anchoring via covalent linkage to the 3′ adenine of the target DNA. Should the polyamide be biotinylated, then precipitation with avidin resins can facilitate the isolation of these fragments for sequencing. Several reports have experimentally demonstrated the ability to enrich these fragments (Figure 5c), and we anticipate Chem-seq to be a versatile platform in the characterization of off-target effects for PIPs. A parallel variation of Chem-seq is COSMIC-seq [58], which is formulated upon the use of psoralen-conjugated PIPs for chromatin crosslinking and enrichment sequencing in conjunction to CSI. This method generates data that infer the location of cognate sites and is primarily focused on the monitoring and assessing of changes in the chromatin states; this method, however, is similar to the aforementioned combination of Bind-n-Seq with whole-genome sequencing in that it also indirectly maps the binding landscape with its motif-based approach.

At its current inception, aside from the need for further experimental optimization, Chem-seq faces a critical challenge with the subsequent data analysis. Those who encounter sequencing data have heard or made similar complaints: while data generation is straightforward, post-acquisition analysis is complicated. This is perhaps one of the reasons why few in the discipline spend significant effort on Chem-seq experiments: processing genomic data can be unforgiving, especially when the toolkits are insufficient. Next-generation sequencing data are noisy in nature; for instance, read duplication, natural or artificial, can potentially account for 30%-70% of the observed sequencing coverage [59]; additionally, the "blacklist" locations in the human genome typically found to produce anomalous and unreliable signal artifacts in sequencing experiments [60] span nearly 9% of the human reference genome (Appendix A), roughly three times the length of protein-coding genes [61]. Unlike proteins, PIP-DNA interactions are restricted to the minor groove; with their short motif recognition, the three-dimensional binding surfaces of PIP-DNA ligands will be smaller than typical protein-DNA ligands. Practically, peak callers written for ChIP-seq experiments will not be ideal in the analysis of Chem-seq data. This includes MACS [62], a popular model-based peak-calling tool ubiquitously used in ChIP-seq applications. Such a tool assumes that the distribution of peaks through the genome follows roughly a Poisson distribution process, and after shifting reads on the opposing strands, peaks are assumed to follow this distribution and be consequently called; with alkylating PIPs, the size of the binding site is essentially down to a single base, which exhibits wholly different binding interactions compared to typical transcription factors. This hypothesis, however, remains to be verified, since the single-base interaction surface located deep in the minor groove will render structural validation an exceedingly difficult task. As a matter of fact, even Anders' original implementation had to utilize a lower enrichment threshold for Chem-seq, involving biotinylated AT7519, and the report was accompanied with generally lower magnitudes at the same sites [57], further highlighting the need for tailored computational tools in Chem-seq post-sequencing analyses.

The above hypothesis is hardly unique; various studies have compared the performance of a number of peak callers with different algorithmic backends in non-ChIP-seq experiments. Koohy and colleagues, for instance, found considerable discrepancies with DNase-seq data [63]; additionally, Poisson-based MACS has also been said to underperform in FAIRE-seq experiments compared to some of the other callers utilizing negative binomial models [64]. Certainly, while sequencing experiments such as ChIP-seq, DNase-seq and even Chem-seq generate the same type of short-read tags, differences in sequencer technology and the chemistry behind the enrichment of nucleotide fragments necessitate that the tools be at a minimum re-tune in order to achieve optimal performance. The differences in the level of background noise between different sequencing methods also cannot be ignored; DNase-seq is highly prone to noise and generates reads that are not strand-specific to have characteristic shifts that are often used to as a marker for positive peaks. Preliminary assessment of sequencing data suggested some similarities between Chem-seq and DNase-seq [65], but there have been no systematic studies to verify this to date.

Alternatives to a Poisson model-based approach include, certainly, the building of a different model. As with the case of ZINBA [64], negative binomial models may be suitable. However, considering the use of Bayesian models in computational biology, for instance, Spyrou and colleagues' proposal [66] that Bayesian hidden Markov models can be used to detect genomic enrichments, increasing the model complexity that likely will improve the prediction outcome. A direct adaptation of the proposed Bayesian approach, which still depends on the presence of somewhat symmetric reads on both forward and reverse strands in situ (as a consequence of the larger protein-DNA-binding surfaces), will nonetheless require modifications, especially when most of the algorithms were developed with analyzing Illumina data in mind, not 100bp or longer, as they are today in the age of semiconductor (Ion Torrent) and long-read (PacBio) sequencers; this inevitably changes the peak distribution in Chem-seq data, such that extensive retooling is still required. With this kind of effort required, it is only imaginable that those in the discipline of PIP research will opt not to traverse down this path.

A tentative solution is to validate peak locations by the seemingly naïve assumption that true positive sites are enriched as the consequence of more affinity enrichment during the experiment, and thus, read enrichment is correlated. Under this assumption, one should be able to infer PIP-binding sites from their relative enrichment in the aligned reads. Based on this line of thinking, an approach is to utilize tools such as diffReps [67] or CRED [68], which, instead of a model-based approach, calculates read enrichments within differential sliding windows to deduce candidate sites. Existing studies utilizing MACS tended to underperform compared to enrichment-based methods at a high margin; for instance, we tested MACS on a CBI-conjugated PIP designed to bind G12D/V mutations of *KRAS* [65] and found that, on default settings, the model generated extremely unrealistic numbers of statistically significant candidates; furthermore, using a set of simulated Ion Torrent reads as the control, MACS also underreported nearly 200 sites compared to CRED [69]. This is not a criticism of MACS as a peak detection program but more of the point that, as complexity in sequencing data increases, no longer can we apply the same *lieu commun* assumptions in data analysis. With that said, enrichment-based approaches should nonetheless be coupled with secondary validations such as additional statistical validations over nonbinding sites of the same motif [65]. It is, however, important to note that, without moving away from motif-based approaches, the performance of enrichment-based peak calling will not fundamentally improve in Chem-seq applications, since the same issues continue to persist.

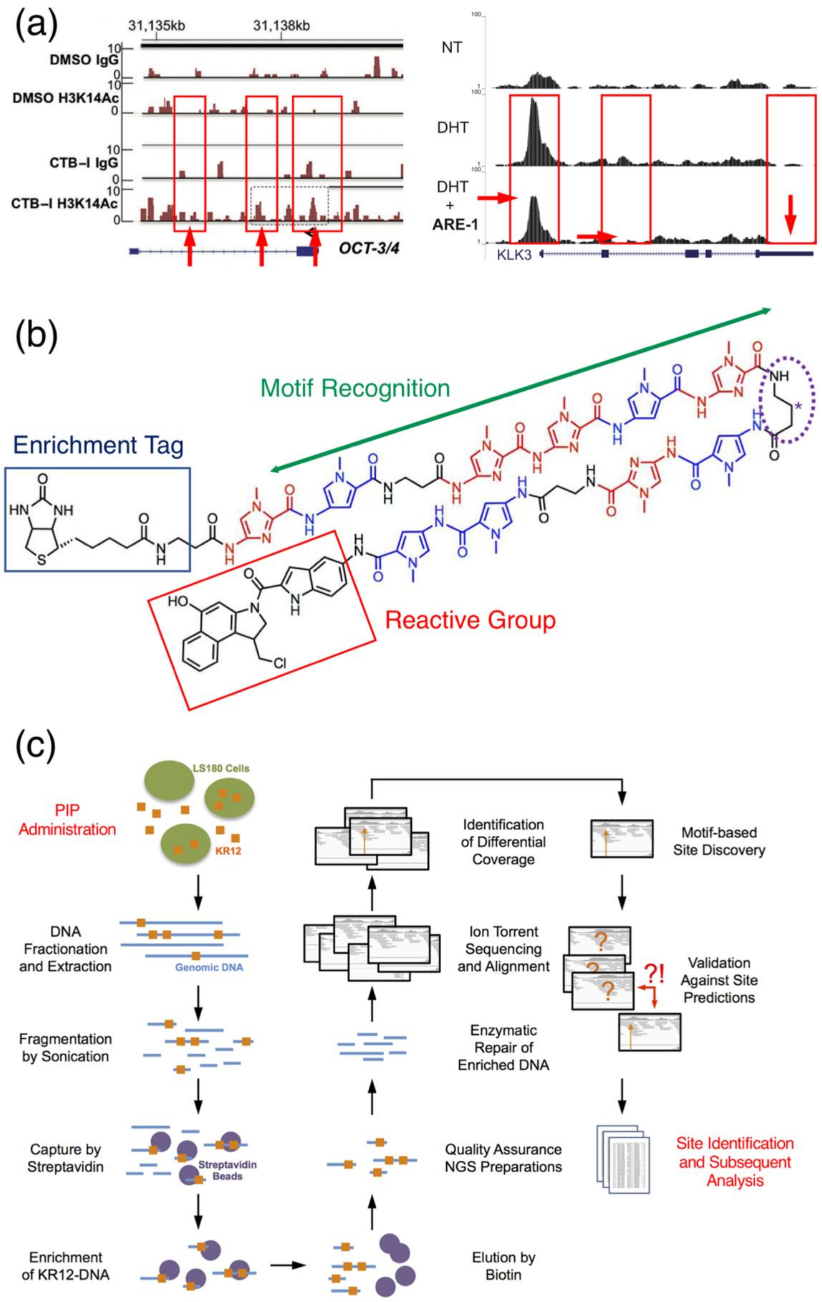
(a)
31,135kb
31,138kb
DMSO IgG
DMSO H3K14Ac
CTB–I IgG
CTB–I H3K14Ac
OCT-3/4
NT
DHT
DHT + ARE-1
KLK3
(b)
Motif Recognition
Enrichment Tag
Reactive Group
(c)
LS180 Cells
PIP Administration
KR12
DNA Fractionation and Extraction
Genomic DNA
Fragmentation by Sonication
Capture by Streptavidin
Streptavidin Beads
Enrichment of KR12-DNA
Elution by Biotin
Quality Assurance NGS Preparations
Enzymatic Repair of Enriched DNA
Ion Torrent Sequencing and Alignment
Identification of Differential Coverage
Motif-based Site Discovery
Validation Against Site Predictions
Site Identification and Subsequent Analysis

Figure 5. Sequencing with PIPs. (**a**) Examples of PIP-assisted ChIP-seq experiments using a CTB, a histone acetyltransferase activator, conjugated PIP in conjunction with ChIP-seq to detect changes in H3K14Ac activity in the promoter region of *OCT-3*/4 (left); another example of coupling dihydrotestosterone (DHT) with a PIP targeting the androgen receptor gene (*AR*) to evaluate changes in the genomic occupancy of *AR* in the promoter region of *KLK3* is provided on the right. *NT* indicates no treatment. Loss, gain or changes in amplitude of the local reads (red arrows and box with solid outlines) are then used to infer PIP genomic interference. Reproduced from [21], © 2015 Wiley (left) and [30], © 2019 Oxford University Press under the license terms of CC BY (right). (**b**) Scheme of a possible configuration of biotinylated PIPs for Chem-seq applications. In these situations, biotin is used as the enrichment tag (solid blue outline), and indole-*seco*-CBI (solid red outline) is the reactive alkylating moiety. The polyamide backbone provides a mean for the molecule to recognize and bind the minor groove of specific genomic DNA sites. Placement of the biotin enrichment tag at the hairpin turn (dotted purple outline labeled with asterisk) has also been proposed and tested [70]. (**c**) A representative Chem-seq workflow utilizing a bifunctionally conjugated PIP (see Figure 4b) to alkylate and enrich genomic-binding sites. Images for (**b**,**c**) were adapted from [65] by the author, © 2019 under the license terms of CC BY.

4. Perspectives

What, then, will entice PIP researchers to head down this road not traveled? Certainly, the most direct method of reducing off-target effects with PIPs is to improve its binding affinity to its target, and approaches aimed at both increasing the recognition motif by conjoining multiple PIPs in tandem or modifying the base chemistry of the heterocycles will improve a polyamide's affinity, but as its size approaches, the nuclear penetration limiting the law of diminishing returns is going to be well in effect. Alternatively, improvements in elucidating such details from expression profiling or Chem-seq data analysis by way of improving the computational analysis may perhaps be more viable. For instance, by leveraging the use of machine-learning methods to "study" Chem-seq peak characteristics (Figure 6a), the peak-calling process may see improved recognition. While Chem-seq peaks may exhibit nonuniform read distributions not fully expressible with the use of a single distribution model, there should nonetheless be key characteristics underlying the read pileup, and ensemble learning methods such as random forests [71] may be well-suited to recognize these feature characteristics. These methods "grow" a multitude of bifurcated decision trees that propagate through the various feature parameters during training, and consensus decisions from these ensemble of decision trees will generate reliable classification results based on the mode of the ensemble. Random forests are notably well-resistant to overfitting and, thus, can hypothetically create versatile predictor models across PIPs with small but generally similar major chemical structures and conjugated functional moieties.

Random forests see a large number of applications in computational biology, such as classification and regression problems of detecting evolutionary events [72], identifying DNA-binding proteins [73] and even predicting drug sensitivity [74]. In biology, where noise is a frequent issue, beyond the typical application of random forests in epidemiological studies for learning behavioral patterns among cohort statistics, this particular type of machine learning has also seen a spike in genomic studies [75]. Additionally, random forests generate deciding factors useful for estimating the relative importance of input parameters in a very natural way of explaining the observed phenomenon. These metrics, e.g., Gini importance scores [76], are often directly proportional to how frequent a particular feature is used to make decisions and can be translated in meaningful ways that directly illuminate the role of a particular biological feature. When coupled with enrichment-based callers, as mentioned in the previous section, machine-learning classifiers can be incorporated as a validation tool that is motif-independent.

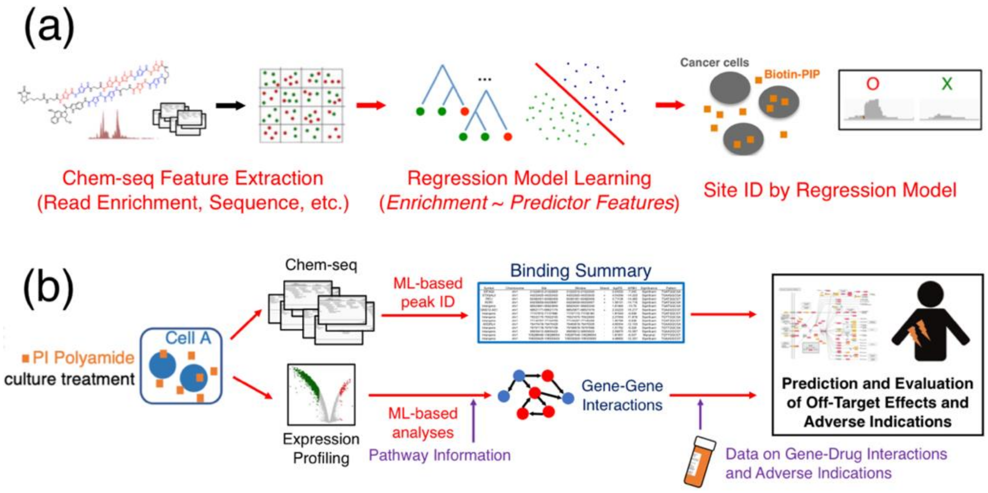
(a)
Cancer cells
Biotin-PIP
O
X
Chem-seq Feature Extraction
(Read Enrichment, Sequence, etc.)
Regression Model Learning
(*Enrichment ~ Predictor Features*)
Site ID by Regression Model
(b)
Chem-seq
ML-based
peak ID
Binding Summary
PI Polyamide
culture treatment
Cell A
Expression
Profiling
ML-based
analyses
Pathway Information
Gene-Gene
Interactions
Prediction and Evaluation
of Off-Target Effects and
Adverse Indications
Data on Gene-Drug Interactions
and Adverse Indications

Figure 6. The use of machine learning to improve PIP off-target analysis from high-throughput data. (**a**) A proposal for using machine-learning methods to improve the site identification process in Chem-seq post-sequencing analysis. (**b**) A scheme of incorporating machine learning (ML) in characterizing genomic binding, as well as expression profiling data, to generate genetic interactions to elucidate the role of off-target effects for candidate PIPs.

We may also consider the use of deep learning, such as neural networks, to improve Chem-seq analysis. Neural networks of multiple layers of decision-making nodes are able to learn features in a dataset and have been used to infer gene expressions, and similar techniques have been applied to reduce noise in ChIP-seq data [77]. The ability to predict gene expressions [78], and epigenetic features such as chromatin structures [79], and enhancer sites [80] theoretically make deep learning also applicable in identifying peaks in Chem-seq data, considering the presence of high-background noise and nonuniform read distributions. Just as sequencing data can be thought of as images, they can be treated as inputs for convolutional neural networks and be processed similarly. For instance, if we recode sequence information as Boolean matrices of positions and features, we can then use layers of perception neurons to learn specific characteristics present in the Chem-seq data to predict the likelihood of a particular alignment being a positive peak.

An inherent problem with neural networks, however, is the cost of computing, especially as the number of neurons increases and the topology of the networks deepens. This problem is further exacerbated as model-training libraries increasingly offload the bulk of computations to graphical processing units (GPU), severely limiting code portability [81] and driving up development costs due to hardware preference. Another hurdle is the lack of well-constructed existing models that can be adapted to improve the learning results. "Transfer learning" is a method of deep learning that seeks to apply pretrained classifier models to a similar constructed problem. Recent advances in artificial intelligence have led to the explosion of image processing and learning models, such as ImageNet, Inception, VGG, etc. [82–84], all of which are extremely valuable in medical applications at identifying the edges of tumor and normal tissues in endoscopy, differentiating physiological features and so on [85]. These prebuilt models, however, may not be as useful for genomics data, which tend to have multiple output classifications, as well as large background noise and, subsequently, bias. A model erring on the side of caution in predicting a marginal peak as noise will tend to have lower false positives and artificially better performance, and in these situations, these models may not be useful at all.

As Frost epiphanies in *The Road Not Taken* [86], "two roads diverged in a wood, and I– I took the one less traveled by, and that has made all the difference," at times, the decision of selecting a different path can change the trajectory of discovery. Minor-groove PIPs have made noticeable strides in the fields of cancer and disease biology, and their binding specificity and relative ease of administration have made them promising candidates in drug discovery; however, the mechanism behind their seeming ability for tumor homing and inhibitory actions remain poorly understood, and despite their specificity, very few have surfaced in clinical trials. Improvements made in the ability to analyze massive parallel-sequencing and expression-profiling data, for instance, by incorporating optimized machine-learning models to denoise Chem-seq data and predictor models for adverse indications (Figure 6b) can potentially aid researchers in understanding what the possible off-target effects may be for a new PIP. These efforts will certainly allow PIPs finally to be funneled into the process of lead development and improve their viability as drug candidates. At this point, unfortunately, few have elected for this road not taken; given the surprising progress of utilizing synthetic heterocycles as minor groove-binding PIPs today, there is something bound to be missed on the other path.

Funding: This work was partly supported by Grant-in-Aids for Scientific Research B and for Young Scientists B from the Japan Society for the Promotion of Science (JP17H03602 to H.N. and JP17K15047 to J.L.), as well as the Japan Agency for Medical Research and Development (AMED; JP18ae0101051h0001, JP17cm0106510, JP17ck0106263 and JP17ck0106356 to H.N. and JP18ck0106422 to H.N. and J.L.).

Acknowledgments: The authors would like to express their deepest gratitude to Paul Horton (Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan) and Tony Kuo (Centre of Biodiversity Genomics, Guelph, Canada) for their generous technical assistance and advice during their former appointments at the Computational Biology Research Center (now part of the Artificial Intelligence Research Center) at the National Institute of Advanced Industrial Science and Technology of Japan. Additionally, the authors would also like to acknowledge former colleagues involved in the writing process, particularly Dr. Hiroyuki Yoda for his help during the initial planning stage.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

Appendix A

The claim that 9% of the human genome consists of "blacklisted regions" prone to producing artificially high signals is based on an estimate of the nonoverlapping length (approximately 274,970,000 bp) of all genomic segments annotated by Boyle et al. (version 2, released Nov. 2018) [87] compared to the length of the hg19 human genome (3,095,677,412 bp, excluding the mitochondrion, alternative haplotypes and random contigs).

References

1. Chang, A.Y.; Dervan, P.B. Strand selective cleavage of DNA by diastereomers of hairpin polyamide-*seco*-CBI conjugates. *J. Am. Chem. Soc.* **2000**, *122*, 4856–4864. [CrossRef]
2. Baird, E.E.; Dervan, P.B. Solid phase synthesis of polyamides containing imidazole and pyrrole amino acids. *J. Am. Chem. Soc.* **1996**, *118*, 6141–6146. [CrossRef]
3. Chenoweth, D.M.; Harki, D.A.; Dervan, P.B. Solution-phase synthesis of pyrrole-imidazole polyamides. *J. Am. Chem. Soc.* **2009**, *131*, 7175–7181. [CrossRef] [PubMed]
4. Dang, C.V.; Reddy, E.P.; Shokat, K.M.; Soucek, L. Drugg the 'undruggable' cancer targets. *Nat. Rev. Cancer* **2017**, *17*, 502–508. [CrossRef] [PubMed]
5. Dayton, J.B.; Piccolo, S.R. Classifying cancer genome aberrations by their mutually exclusive effects on transcription. *BMC Med. Genomics* **2017**, *10* (Suppl. 4), 66. [CrossRef] [PubMed]
6. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **2012**, *489*, 519–525. [CrossRef]
7. Rudmann, D.G. On-target and off-target-based toxicologic effects. *Toxicol. Pathol.* **2012**, *41*, 310–314. [CrossRef]
8. Kawamoto, Y.; Sasaki, A.; Hashiya, K.; Ide, S.; Bando, T.; Maeshima, K.; Sugiyama, H. Tandem trimer pyrrole-imidazole polyamide probes targeting 18 base pairs in human telomere sequences. *Chem. Sci.* **2015**, *6*, 2307–2312. [CrossRef]
9. Kawamoto, Y.; Sakai, A.; Chandran, A.; Hashiya, K.; Ide, S.; Bando, T.; Maeshima, K.; Sugiyama, H. Targeting 24 bp within telomere repeat sequences with tandem tetramer pyrrole-imidazole polyamide probes. *J. Am. Chem. Soc.* **2016**, *138*, 14100–14107. [CrossRef]
10. Paine, P.L.; Moore, L.C.; Horowitz, S.B. Nuclear envelope permeability. *Nature* **1975**, *254*, 109–114. [CrossRef]
11. Keminer, O.; Peters, R. Permeability of single nuclear pores. *Biophys. J.* **1999**, *77*, 217–228. [CrossRef]
12. Mou, H.; Smith, J.L.; Peng, L.; Yin, H.; Moore, J.; Zhang, X.O.; Song, C.Q.; Sheel, A.; Wu, Q.; Ozata, D.M.; et al. CRISPR/Cas9-mediated genome editing induces exon skipping by alternative splicing or exon deletion. *Genome Biol.* **2017**, *18*, 108. [CrossRef] [PubMed]
13. Vitaku, E.; Smith, D.T.; Njardarson, J.T. Analysis of the structural diversity, substation patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem.* **2014**, *57*, 10257–10274. [CrossRef] [PubMed]
14. Gibson, D. Drug-DNA interactions and novel drug design. *Pharm. J.* **2002**, *2*, 275–276. [CrossRef] [PubMed]
15. Muzikar, K.A.; Meier, J.L.; Gubler, D.A.; Raskatov, J.A.; Dervan, P.B. Expanding the repertoire of natural product-inspired ring pairs for molecular recognition of DNA. *Org. Lett.* **2011**, *13*, 5612–5615. [CrossRef]
16. Trauger, J.W.; Baird, E.E.; Dervan, P.B. Extended hairpin polyamide motif for sequence-specific recognition in the minor groove of DNA. *Chem. Biol.* **1996**, *3*, 369–377. [CrossRef]
17. Xu, L.; Wang, W.; Gotte, D.; Yang, F.; Hare, A.A.; Welch, T.R.; Li, B.C.; Shin, J.H.; Chong, J.; Strathern, J.N.; et al. RNA polymerase II senses obstruction in the DNA minor groove via a conserved sensor motif. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12426–12431. [CrossRef]
18. Philips, B.J.; Chang, A.Y.; Dervan, P.B.; Beerman, T.A. DNA damage effects of a polyamide-CBI conjugate in SV40 virions. *Mol. Pharmacol.* **2005**, *67*, 877–882. [CrossRef]
19. Hang, Z.; Gao, J.; Chen, Z.; Duan, S.; Li, C.; Qiao, R. Double-strand cleavage of DNA by a polyamide-phenazine-di-N-oxide conjugate. *Bioorg. Med. Chem. Lett.* **2018**, *28*, 284–288.

20. Pandian, G.N.; Sato, S.; Anandhakumar, C.; Taniguchi, J.; Takashima, K.; Syed, J.; Han, L.; Saha, A.; Bando, T.; Nagase, H.; et al. Identification of a small molecule that turns on the pluripotency gene circuitry in human fibroblasts. *ACS Chem. Biol.* **2014**, *9*, 2729–2736. [CrossRef]
21. Han, L.; Pandian, G.N.; Chandran, A.; Sato, S.; Taniguchi, J.; Kashiwazaki, G.; Sawatani, Y.; Hashiya, K.; Bando, T.; Xu, Y.; et al. A synthetic DNA-binding domain guides distinct chromatin-modifying small molecules to activate an identical gene network. *Angew. Chem. Int. Ed.* **2015**, *54*, 8700–8703. [CrossRef] [PubMed]
22. Inoue, T.; Shimozato, O.; Matsuo, N.; Mori, Y.; Shinozaki, Y.; Lin, J.; Watanabe, T.; Takatori, A.; Koshikawa, N.; Ozaki, T.; et al. Hydrophobic structure of hairpin ten-ring pyrrole-imidazole polyamides enhances tumor tissue accumulation/retention in vivo. *Bioorg. Med. Chem.* **2018**, *26*, 2337–2344. [CrossRef] [PubMed]
23. Harki, D.A.; Satyamurthy, N.; Stout, D.B.; Phelps, M.E.; Dervan, P.B. In vivo imaging of pyrrole-imidazole polyamides with positron emission tomography. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13039–13044. [CrossRef] [PubMed]
24. Kojima, T.; Wang, X.; Fujiwara, K.; Osaka, S.; Yoshida, Y.; Osaka, E.; Taniguchi, M.; Ueno, T.; Fukuda, N.; Soma, M.; et al. Inhibition of human osteosarcoma cell migration and invasion by a gene silencer, pyrrole-imidazole polyamide, targeted at the human MMP9 NF-kB binding site. *Biol. Pharma. Bull.* **2014**, *37*, 1460–1465. [CrossRef]
25. Hiraoka, K.; Inoue, T.; Taylor, R.D.; Watanabe, T.; Koshikawa, N.; Yoda, H.; Shinohara, K.I.; Takatori, A.; Sugimoto, H.; Maru, Y.; et al. Inhibition of KRAS codon 12 mutants using a novel DNA-alkylating pyrrole-imidazole polyamide conjugate. *Nat. Commun.* **2015**, *6*, 6706. [CrossRef]
26. Morita, K.; Suzuki, K.; Maeda, S.; Matsuo, A.; Mitsuda, Y.; Tokushige, C.; Kashiwazaki, G.; Taniguchi, J.; Maeda, R.; Noura, M.; et al. Genetic regulation of the RUNX transcription factor family has antitumor effects. *J. Clin. Investig.* **2017**, *127*, 2815–2828. [CrossRef]
27. Yoda, H.; Inoue, T.; Shinozaki, Y.; Lin, J.; Watanabe, T.; Koshikawa, N.; Takatori, A.; Nagase, H. Direct targeting of MYCN gene amplification by site-specific DNA alkylation in neuroblastoma. *Cancer Res.* **2019**, *79*, 830–840. [CrossRef]
28. Szablowski, J.O.; Raskatov, J.A.; Dervan, P.B. An HRE-binding Py-Im polyamide impairs hypoxic signaling in tumors. *Mol. Cancer Ther.* **2016**, *15*, 608–617. [CrossRef]
29. Hargrove, A.E.; Martinez, T.F.; Hare, A.A.; Kurmis, A.A.; Phillips, J.W.; Sud, S.; Pienta, K.J.; Dervan, P.B. Tumor repression of VCaP xenografts by a pyrrole-imidazole polyamide. *PLoS ONE* **2015**, *10*, e0143161. [CrossRef]
30. Kurmis, A.A.; Dervan, P.B. Sequence specific suppression of androgen receptor–DNA binding in vivo by a Py-Im polyamide. *Nucleic Acids Res.* **2019**, *47*, 3828–3835. [CrossRef]
31. Pommier, Y.; Leo, E.; Zhang, H.; Marchand, C. DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem. Biol.* **2010**, *7*, 421–433. [CrossRef] [PubMed]
32. Raskatov, J.A.; Meier, J.L.; Puckett, J.W.; Yang, F.; Ramakrishnan, P.; Dervan, P.B. Modulation of NF-kB-dependent gene transcription using programmable DNA minor groove binders. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1023–1028. [CrossRef] [PubMed]
33. Kang, J.J.S.; Dervan, P.B. A sequence-specific DNA binding small molecule triggers the release of immunogenic signals and phagocytosis in a model of B-cell lymphoma. *Q. Rev. Biophys.* **2015**, *48*, 453–464. [CrossRef] [PubMed]
34. Igarashi, J.; Fukuda, N.; Inoue, T.; Nakai, S.; Saito, K.; Fujiwara, K.; Matsuda, H.; Ueno, T.; Matsumoto, Y.; Watanabe, T.; et al. Preclinical study of novel gene silencer pyrrole-imidazole polyamide targeting human TGF-β1 promoter for hypertrophic scars in a common marmoset primate model. *PLoS ONE* **2015**, *10*, e0125295. [CrossRef]
35. Van Norman, G.A. Phase II trials in drug development and adaptive trial design. *JACC Basic Transl. Sci.* **2019**, *4*, 428–437. [CrossRef]
36. Broder, S. The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antivir. Res.* **2010**, *85*, 1–18. [CrossRef]
37. Terrett, N.K.; Bell, A.S.; Brown, D.; Elllis, P. Sildenafil (Viagra), a potent and selective inhibitor of Type 5 cGMP phosphodiesterase with utility for the treatment of male erectile dysfunction. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 1819–1824. [CrossRef]

38. Jespersen, C.; Soragni, E.; Chou, C.J.; Arora, P.S.; Dervan, P.B.; Gottesfeld, J.M. Chromatin structure determines accessibility of a hairpin polyamide-chlorambucil conjugate at histone H4 genes in pancreatic cancer cells. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4068–4071. [CrossRef]
39. Kitagawa, Y.; Okumura, K.; Watanabe, T.; Tsukamoto, K.; Kitano, S.; Nankinzan, R.; Suzuki, T.; Hara, T.; Soda, H.; Denda, T.; et al. Enrichment technique to allow early detection and monitor emergence of KRAS mutation in response to treatment. *Sci. Rep.* **2019**, *9*, 11346. [CrossRef]
40. Han, Y.W.; Kashiwazaki, G.; Morinaga, H.; Matsumoto, T.; Hashiya, K.; Bando, T.; Harada, Y.; Sugiyama, H. Effect of single pyrrole replacement with b-alanine on DNA binding affinity and sequence specificity of hairpin pyrrole/imidazole polyamides targeting 5′-GCGC-3′. *Bioorg. Med. Chem.* **2013**, *21*, 5436–5441. [CrossRef]
41. Lacy, E.R.; Cox, K.K.; Wilson, W.D.; Lee, M. Recognition of T*G mismatched base pairs in DNA by stacked imidazole-containing polyamides: Surface plasmon resonance and circular dichroism studies. *Nucleic Acid Res.* **2002**, *30*, 1834–1841. [CrossRef] [PubMed]
42. Qiao, H.; Ma, C.; Zhang, X.; Jing, X.; Li, C.; Zhao, Y. Insight into DNA minor groove unspecific binding of pyrrole polyamide. *Bioconjugate Chem.* **2015**, *26*, 2054–2061. [CrossRef] [PubMed]
43. Muller, S.; Paulus, J.; Mattay, J.; Ihmels, H.; Dodero, V.I.; Sewald, N. Photocontrolled DNA minor groove interactions of imidazole/pyrrole polyamides. *Beilstein J. Org. Chem.* **2020**, *16*, 60–70. [CrossRef] [PubMed]
44. Svaren, J.; Klebanow, E.; Sealy, L.; Chalkley, R. Analysis of the competition between nucleosome formation and transcription factor binding. *J. Biol. Chem.* **1994**, *269*, 9335–9344.
45. Raskatov, J.A.; Szablowski, J.O.; Dervan, P.B. Tumor xenograft uptake of a pyrrole–imidazole (Py-Im) polyamide varies as a function of cell line grafted. *J. Med. Chem.* **2014**, *57*, 8471–8476. [CrossRef]
46. Hargrove, A.E.; Raskatov, J.A.; Meier, J.L.; Montgomery, D.C.; Dervan, P.B. Characterization and solubilization of pyrrole–imidazole polyamide aggregates. *J. Med. Chem.* **2012**, *55*, 5425–5432. [CrossRef]
47. Yang, F.; Nickols, N.G.; Li, B.C.; Szablowski, J.O.; Hamilton, S.R.; Meier, J.L.; Wang, C.M.; Dervan, P.B. Animal toxicity of hairpin pyrrole-imidazole polyamides varies with the turn unit. *J. Med. Chem.* **2013**, *56*, 7449–7457. [CrossRef]
48. Mishra, R.; Watanabe, T.; Kimura, M.T.; Koshikawa, N.; Ikeda, M.; Uekusa, S.; Kawashima, H.; Wang, X.; Igarashi, J.; Choudhury, D.; et al. Identification of a novel E-box binding pyrrole-imidazole polyamide inhibiting *MYC*-driven cell proliferation. *Cancer Sci.* **2015**, *106*, 421–429. [CrossRef]
49. Lin, J.; Krishnamurthy, S.; Yoda, H.; Shinozaki, Y.; Watanabe, T.; Koshikawa, N.; Takatori, A.; Horton, P.; Nagase, H. Estimating genome-wide off-target effects of pyrrole-imidazole polyamide binding by a pathway-based expression profiling approach. *PLoS ONE* **2019**, *14*, e0215247. [CrossRef]
50. Yera, E.R.; Cleves, A.E.; Jain, A.N. Chemical structural novelty: On-targets and off-targets. *J. Med. Chem.* **2011**, *54*, 6771–6785. [CrossRef]
51. Zykovich, A.; Korf, I.; Segal, D.J. Bind-n-Seq: High-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res.* **2009**, *37*, e151. [CrossRef] [PubMed]
52. Kang, J.S.; Meier, J.L.; Dervan, P.B. Design of sequence-specific DNA binding molecules for DNA methyltransferase inhibition. *J. Am. Chem. Soc.* **2014**, *136*, 3687–3694. [CrossRef]
53. Chandran, A.; Syed, J.; Taylor, R.D.; Kashiwazaki, G.; Sato, S.; Hashiya, K.; Bando, T.; Sugiyama, H. Deciphering the genomic targets of alkylating polyamide conjugates using high-throughput sequencing. *Nucleic Acids Res.* **2016**, *44*, 4014–4024. [CrossRef] [PubMed]
54. Kashiwazaki, G.; Chandran, A.; Asamitsu, S.; Kawase, T.; Kawamoto, Y.; Hashiya, K.; Bando, T.; Sugiyama, H. Comparative analysis of DNA-binding selectivity of hairpin and cyclic pyrrole-imidazole polyamides based on next-generation sequencing. *ChemBioChem* **2016**, *17*, 1752–1758. [CrossRef] [PubMed]
55. Carlson, C.D.; Warren, C.L.; Hauschild, K.E.; Ozers, M.S.; Qadir, N.; Bhimsaria, D.; Lee, Y.; Cerrina, F.; Ansari, A.Z. Specificity landscapes of DNA binding molecules elucidate biological function. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4544–4549. [CrossRef] [PubMed]
56. Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y.C.; Laslo, P.; Cheng, J.X.; Murre, C.; Singh, H.; Glass, C.K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **2010**, *38*, 576–589. [CrossRef]
57. Anders, L.; Guenther, M.G.; Qi, J.; Fan, Z.P.; Marineau, J.J.; Rahl, P.B.; Loven, J.; Sigova, A.A.; Smith, W.B.; Lee, T.I.; et al. Genome-wide localization of small molecules. *Nat. Biotechnol.* **2014**, *32*, 92–96. [CrossRef]

58. Erwin, G.S.; Bhimsaria, D.B.; Eguchi, A.; Ansari, A.Z. Mapping polyamide-DNA interactions in human cells reveals a new design strategy for effective targeting of genomic sites. *Angew. Chem. Int. Ed.* **2014**, *53*, 10124–10128. [CrossRef]
59. Zhou, W.; Chen, T.; Zhao, H.; Eterovic, A.K.; Meric-Bernstam, F.; Mills, G.B.; Chen, K. Bias from removing read duplication in ultra-deep sequencing experiments. *Bioinformatics* **2014**, *30*, 1073–1080. [CrossRef]
60. Amemiya, H.M.; Kundaje, A.; Boyle, A.P. The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Rep.* **2019**, *9*, 9354. [CrossRef]
61. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74. [CrossRef] [PubMed]
62. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **2008**, *9*, R137. [CrossRef] [PubMed]
63. Koohy, H.; Down, T.A.; Spivakov, M.; Hubbard, T. A comparison of peak callers used for DNase-seq data. *PLoS ONE* **2014**, *9*, e96303. [CrossRef] [PubMed]
64. Rashid, N.U.; Giresi, P.G.; Ibrahim, J.G.; Sun, W.; Lieb, J.D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* **2011**, *12*, R67. [CrossRef] [PubMed]
65. Lin, J.; Hiraoka, K.; Watanabe, T.; Kuo, T.; Shinozaki, Y.; Takatori, A.; Koshikawa, N.; Chandran, A.; Otsuki, J.; Sugiyama, H.; et al. Identification of binding targets of a pyrrole-imidazole polyamide KR12 in the LS180 colorectal cancer genome. *PLoS ONE* **2016**, *11*, e0165581. [CrossRef]
66. Spyrou, C.; Stark, R.; Lynch, A.G.; Tavare, S. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinform.* **2009**, *10*, 299. [CrossRef]
67. Shen, L.; Shao, N.Y.; Liu, X.; Maze, I.; Feng, J.; Nestler, E.J. diffReps: Detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE* **2013**, *8*, e65598. [CrossRef]
68. Lin, J.; Kuo, T.; Horton, P.; Nagase, H. CRED: A rapid peak caller for Chem-seq data. *J. Open Source Softw.* **2019**, *4*, 1423. [CrossRef]
69. Lin, J. Chem-Seq Read Enrichment Discovery: CRED (See *Example* Section under README.md). Available online: Github.com/jlincbio/cred/ (accessed on 30 January 2020).
70. Kashiwazaki, G.; Maeda, R.; Kawase, T.; Hashiya, K.; Bando, T.; Sugiyama, H. Evaluation of alkylating pyrrole-imidazole polyamide conjugates by a novel method for high-throughput sequencer. *Bioorg. Med. Chem.* **2018**, *26*, 1–7. [CrossRef]
71. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
72. Kechris, K.J.; Lin, J.C.; Bickel, P.J.; Glazer, A.N. Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9584–9589. [CrossRef] [PubMed]
73. Nimrod, G.; Szilagyi, A.; Leslie, C.; Ben-Tal, N. Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. *J. Mol. Biol.* **2009**, *387*, 1040–1053. [CrossRef] [PubMed]
74. Riddick, G.; Song, H.; Ahn, S.; Walling, J.; Borges-Rivera, D.; Zhang, W.; Fine, H.A. Predicting in vitro drug sensitivity using random forests. *Bioinformatics* **2011**, *27*, 220–224. [CrossRef] [PubMed]
75. Chen, X.; Ishwaran, H. Random forests for genomics data analysis. *Genomics* **2012**, *99*, 323–329. [CrossRef] [PubMed]
76. Nembrini, S.; Konig, I.R.; Wright, M.N. The revival of the Gini importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef]
77. Koh, P.W.; Pierson, E.; Kundaje, A. Denoising genome-wide histone ChIP-seq with convolutional neural networks. *Bioinformatics* **2017**, *33*, i225–i233. [CrossRef]
78. Cuperus, J.T.; Groves, B.; Kuchina, A.; Rosenberg, A.B.; Jojic, N.; Fields, S.; Seelig, G. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* **2017**, *27*, 2015–2024. [CrossRef]
79. Farre, P.; Heurteau, A.; Cuvier, O.; Emberly, E. Dense neural networks for predicting chromatin conformation. *BMC Bioinform.* **2018**, *19*, 372. [CrossRef]
80. Kim, S.G.; Harwani, M.; Grama, A.; Chaterji, S. EP-DNN: A deep neural network-based global enhancer prediction algorithm. *Sci. Rep.* **2016**, *6*, 38433. [CrossRef]

81. Zhang, Y.; Sinclair, M.; Chien, A.A. Improving Performance Portability in OpenCL Programs. In *International Supercomputing Conference Lecture Notes in Computer Science*; Kunkel, J.M., Ludwig, T., Meuer, H.W., Eds.; Springer-Verlag Berlin: Heidelberg, Germany, 2013; Volume 7905, pp. 136–150.
82. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
83. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z.B. Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
84. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, California, USA, 7–9 May 2015.
85. Ribeiro, E.; Uhl, A.; Wimmer, G.; Hafner, M. Exploring deep learning and transfer learning for colonic polyp classification. *Comput. Math. Methods Med.* **2016**, *2016*, 6584725. [CrossRef] [PubMed]
86. Frost, R. The Road Not Taken. Available online: www.poetryfoundation.org/poems/44272/the-road-not-taken (accessed on 14 February 2020).
87. Boyle, A.P. Application for Making ENCODE Blacklists. Available online: Github.com/Boyle-Lab/Blacklist (accessed on 10 March 2020).