Chapter 6

Future perspectives

With the advent of high-throughput DNA sequencing technology, somatic alterations in cancer genomes are now identified at base-pair resolution in ever-expanding patient cohorts across a wide variety of histological subtypes. In contrast to the well-studied catalogues of single nucleotide variants, comprehensive studies of somatic rearrangement have lagged in development, impeded by the intrinsic complexity of their irregular and multifaceted structural forms. Consequently, the cancer genomics field lacks a robust and well-founded methodology for systematic SV specification, visualisation, and annotation.

The main aims of this thesis were to capitalise on a newly collated WGS dataset of somatic SV calls in 2559 cancer samples in order to: survey the diverse panorama of cancer rearrangement in different cell types; analyse signatures of SV form, location, and prevalence; and define a consistent framework for understanding and reporting genome rearrangement to advance the capabilities of future projects. Building on a recently developed classification scheme to identify the precise structure of individual breakpoint junctions and separate out complex clusters, I described the pan-cancer SV landscape of structural features (Chapter 2), genome property associations (Chapter 3), co-occurrence patterns (Chapter 4), and complex events (Chapter 5). To conclude, I highlight opportunities for further research and development with a focus on: algorithms and technology for SV detection (Section 6.1); the need for complete SV classification tools (Section 6.2); open questions regarding SV signature analysis (Section 6.3); and, finally, discovery and annotation of key functional consequences, with the ultimate goal of pinpointing relevant SV drivers of the cancer phenotype in a clinical setting (Section 6.4).

6.1 Identifying somatic genome rearrangement

In the PCAWG dataset used in this thesis, structural variants were identified by discordant and split paired-end sequences from the short-read Illumina Hi-Seq platform. Taking the intersection of SV calls from four different algorithms, the PCAWG consortium aimed to report high-confidence somatic events with congruent copy number support from read depth evidence. Without orthogonal technologies for SV detection, the specificity and sensitivity of this dataset is unknown. In future projects analysing cancer rearrangement, other biotechnology platforms may supplement or supersede the current Illumina pipelines to: validate SV calls with independent data; capture previously unmapped SVs in longer repeat regions; find variation in the 50 bp–1 kb range mostly overlooked by short-read sequencing; and phase BPJ to the same or different derivative chromosomes. New technologies with established benefits for SV detection (germline or somatic) include linked-read sequencing (Greer et al., 2017; Xia et al., 2017), long-read sequencing (Nattestad et al., 2017; Merker et al., 2018), optical mapping (Chan et al., 2017; Jaratlerdsiri et al., 2017), and Hi-C chromosome conformation assays (Harewood et al., 2017). A combination of approaches will yield the richest portrait of rearrangement (Chaisson et al., 2017), subject to comprehensive algorithmic development for integrating disparate lines of evidence within multi-platform datasets.

Although technological developments may eventually render short-read sequencing obsolete, the short to medium term prospects for SV analysis in large patient cohorts is still largely dominated by Illumina data in the legacy repositories of TCGA and ICGC, as well as ongoing sequencing projects by Genomics England (2017) and other initiatives. As such, there remains considerable value in continuing to improve variant calling pipelines for short-read WGS data. Ideally, SV events and CN segmentation would be jointly estimated by one inclusive algorithm aiming to uphold the logical expectation of higher CN states on the read-group side of every genuine breakpoint. The sub-clonal CN calls used in this thesis were obviously inaccurate in a sizeable minority of cases (Section 5.3), leaving an unmet demand for reliable estimation of non-integer CN states in complex and heterogeneous cancer cell populations. Sv detection and CN estimation in the soma is further confounded by germline polymorphism. Cn segmentation may benefit from explicit modelling of the germline SV states found in the matched normal sample, possibly using catalogues of common, population-matched SV inheritance (Sudmant et al., 2015) to better separate the germline events from the cancer-specific alterations.

6.2 Classifying breakpoint junctions

The major insights contributed by this thesis were facilitated by a novel BPJ classification scheme described in Sections 2.1.3 and 2.2.2. Where cancer rearrangement studies were previously limited to a handful of basic SV classes defined by BPJ orientation with one or two additional caveats, I was instead able to leverage twenty well-reasoned SV classes, including local 2-jump subtypes and long looping events of chromoplexy or templated insertion. All downstream investigation benefited from this detailed codification of individual breakpoints, empowering meaningful stratification within every SV property analysis to avoid massive confounding from heterogeneous phenomena. Notwithstanding this advancement, BPJ clustering and taxonomy remain deeply challenging tasks, with over half of all PCAWG breakpoints unexplained by the current system.

In the existing pipeline, BPJ are first clustered into groups with closer than expected proximity given sample-specific SV rates, and then adjacent breakpoints are partitioned into footprints labelled by their break orientation pattern. The final event classification depends on these footprint motifs and the connecting BPJ, shelving all cryptic configurations to a complex unexplained bin. The output depends heavily on the initial BPJ clustering, and it this clustering step which provides the first opportunity for improvement.

The limitations of the current approach (Section 2.1.3) include: the failure to account for BPJ interrelation in loops across multiple loci; the inability to separate distinct clusters connected by an unrelated BPJ; and the dependence on BPJ orientation frequencies oblivious to the broader structural context. I attempted to overcome some of these limitations with an alternative clustering method on the complex unexplained fraction using node-edge graph models (Section 5.1.1). However, in its present implementation, my graph method is also compromised by a reliance on arbitrary thresholds for node partitioning and component merging/separation. Ideally, the next generation of BPJ clustering methods will address these shortcomings in a formal probabilistic framework conditioning on the sample-specific SV composition. I propose that an iterative approach may offer the optimal solution—clustering and classifying by turns until the updates converge on a final stable solution. For example, if a sample has 250 foldback-type BPJ ($\langle ++ \rangle$ or $\langle -- \rangle$), and two such junctions fall within one or two megabases of each other, an initial cluster partition might separate these BPJ into independent events given the high overall rate of this junction class. However, if the subsequent classification step estimates that 220 of these

BPJ are actually explained by one chromothripsis event, the purported rate of isolated foldback would drastically reduce, causing the next cluster iteration to group the two inverting BPJ in one related event such as a dup-trp-dup or small BFB cluster (depending on the orientation). In an iterative framework, the clustering procedure could even account for the estimated location distribution of different event classifications, such as independent tandem duplications enriched in early replicating DNA, and independent deletions enriched in late replicating DNA (especially fragile sites). Sub-clonality provides another line of evidence informing cluster estimation (Cmero et al., 2017), assuming that high-confidence sub-clonal BPJ should only cluster with SV in the same approximate cell fraction. As 'third' (and 'fourth') generation sequencing becomes more ubiquitous, additional phasing information may greatly reduce the ambiguity of SV patterns along homologous chromosomes and/or in different cell fractions.

Given a particular partition of a sample's BPJ terrain, the next logical step is classification of the separated clusters, assuming they are generated by independent (or at least punctuated) rearrangement events. The current classification scheme (Section 2.1.3) is limited to isolated footprints in simple combinations, augmented with a library of possible overlaps to dissect a fraction of those convoluted clusters up to three or four BPJ. From my exploratory analysis of the complex unexplained SV in Chapter 5, I established that a different BPJ clustering scheme may recover additional SV events conforming to simple definitions, and that many more templated insertion and chromoplexy events would be recovered by extending the classification scheme to adjacent and overlapping footprint motifs. Furthermore, the current library of theoretical overlap structures does not account for templated insertion or local 2-jump events, and so upgrading this reference library may readily yield automatic classifications for another tranche of SV clusters.

These avenues for refining the current classification procedure are ultimately limited to small and relatively simple events, as larger clusters rapidly approach a unique parameter space that cannot possibly be afforded individual categories by specific BPJ configurations. At some point, SV classification strategies must transition from bottom-up to top-down, such that complex SV clusters are characterised by their overall feature profile, linking the total formation—where possible—to compatible underlying mechanisms such as chromothripsis, chromoplexy, chromoanasynthesis, and so forth. A top-down view is also more robust to false negative and/or false positive contamination; problems not accommodated by the simple SV classifier assuming complete BPJ information.

An outstanding question is how best to summarise the characteristic attributes of complex SV events in order to generate taxonomical divisions with proven correspondence to the underlying rearrangement mechanism. Thus, future research could consider: *de novo* event clustering using distance measures between independent SV; fixed classification rules trained on clear examples of canonical mechanisms; and computer simulations of genome rearrangement under a range of mechanistic models applied in varying combinations. In any case, the distinctions between different phenomena must be measured via a raft of summary statistics to capture relevant aspects of copy number, orientation, and connectivity. Experimental systems which generate complex SV events via known pathways of breakage and repair may provide additional validation and guidance in optimising these efforts (Meier et al., 2014; Maciejowski et al., 2015; Mardin et al., 2015; Zhang et al., 2015).

Given the importance of somatic rearrangement in cancer biology—and the role of similarly complex germline SVs in developmental disorders (Heesch et al., 2014; Collins et al., 2017)—complete SV specification tools are in high demand for research and clinical use, and must be regarded a major priority of bioinformatic development in the next few years.

6.3 Signatures of mutational process

As a valuable window into cancer aetiology and DNA dynamics, the mutational signatures imparted by different underlying processes are estimated by co-occurrence pattern matching across cancer sample cohorts. Throughout Chapter 4, I discussed my future proposals for extending the current signature paradigm, with particular attention to the hierarchical Dirichlet process. Here, I briefly highlight some open questions in relation to SV signatures in particular. In the abiding signature model, genome alteration classes are tallied as independent events in discrete, unordered categories. This may be a partially false premise in the structural variant realm, with events spanning a wide spectrum of size and complexity without neatly dividing into independent categories of comparable scale. It remains unclear how large, rare events like chromothripsis and chromoplexy should be compared against small, common SV like deletion and tandem duplication. Furthermore, the relevant features of size, microhomology, and replication timing skew, more naturally suit a signature framework of distributions over separate variables rather than discrete categorical observations. Regardless of the model, another frontier for signature research is the integration of SNV, indel, and SV alteration classes within one overarching analysis, possibly using hierarchical models to share information across disparate data types reflecting a shared underlying condition such as HR deficiency or UV radiation.

6.4 Functional consequences of rearrangement

The investigations in this thesis focused mainly on the patterns and properties of somatic rearrangement, irrespective of their functional import as passenger or driver genome alterations. In this section, I discuss the prospects for annotation and selection analysis of functional consequences, as informed by the SV landscape surveyed throughout this work.

6.4.1 Annotation

As reviewed in Section 1.5, one rearrangement event may impart several gene-altering effects, including gene disruption or fusion across breakpoint junctions, gene dosage changes within the span of a SV footprint, and ectopic gene-enhancer regulation within merged or neo-TAD structures. At present, there are no available tools to annotate the full consequence spectrum of BPJ clusters in varying configurations.

For simple SV between two genome positions, it would be feasible to construct a complete atlas of gene-level consequences in the two-dimensional space of all possible events. Figure 6.1 outlines a proposed design for partitioning the space of all possible deletions or tandem duplications along a chromosome into functional consequences for one particular gene of interest. In theory, a simple rule set could construct a similar map for every gene, with any observed event easily annotated by position look-ups across the atlas of relevant maps. Although this construct may seem more convoluted than on-the-fly calculations for individually observed events, the annotation atlas has useful implications for recurrence-based driver analysis, as discussed in the following section.

For more complex SV events spanning multiple genome loci, it is impractical to calculate a full atlas of functional consequences for all possible structures. Instead, the individually observed BPJ and CN profiles could be parsed for likely fusions and dosage change, with ectopic enhancer contacts predicted from TAD boundary placement along likely derivatives. As adjacent BPJ may have

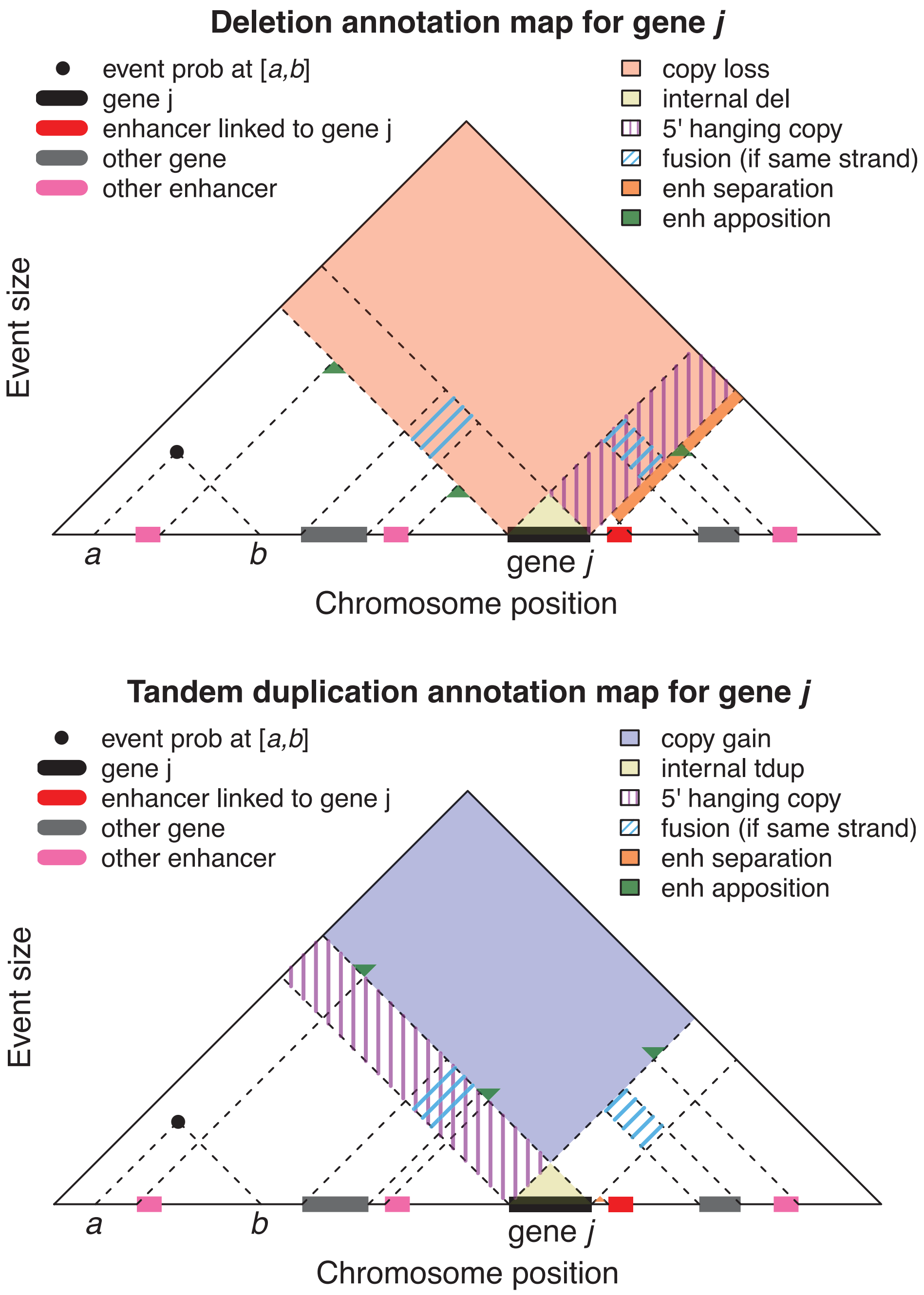

Figure 6.1: Schematic annotation maps of the functional consequences at gene j (assuming a + strand gene) imparted by all possible deletions (upper) or tandem duplications (lower) along the chromosome, with every point in the triangle representing a possible BPJ between two perpendicular points (like a and b).

uncertain phasing, some annotations (particularly enhancer apposition) may best be reported as probabilistic possibilities given the sample-specific likelihood of adjacent breakpoints occurring by chance on different chromosomes.

Existing knowledge banks of established cancer genes can be utilised to highlight putative driver SV events. For example, Notta et al. (2016) annotated pancreatic cancer rearrangements with simultaneous knockout of several canonical cancer genes. Aside from highlighting alterations to known oncogenes and tumour suppressors, it would also be beneficial to assess which of the many other functional consequences have relevance to cancer progression. Although hundreds of genes have already been labelled with known cancer effects, many more may yet be found, with one recent estimate suggesting half of all coding SNV drivers occur outside known cancer genes (Martincorena et al., 2017).

6.4.2 Driver discovery

So far, SV driver discovery efforts have focussed on: foci of recurrent copy gain or copy loss (Beroukhim et al., 2010; Mermel et al., 2011); enhancer-hijacking (Weischenfeldt et al., 2017); and one or two dimensional breakpoint recurrence, agnostic to BPJ classification (Wala et al., 2017a). As discussed in Chapter 3, different SV classes have markedly different formation rates across the genome, and, therefore, recurrence-based driver discovery should ideally account for structure-specific (and tissue-specific) background distribution estimates (before selection), in concert with sample-specific SV class exposures. Additionally, it would be preferable to integrate multiple effects—dosage, disruption, fusion, and regulation—to maximise available evidence for positive selection at the level of individual genes.

To this end, I return to the annotation map concept illustrated in Figure 6.1. For a given set of observed annotations, the question arises: which of these functional effects has occurred significantly more or less often than expected in the absence of selection? If we could determine the background probability of every possible SV event—that is, the probability at each point in the annotation map—then the expected rate of each annotated consequence before selection is the summation of event probabilities within the relevant partition. In this way, effects can be integrated across disparate SV classes while upholding the class-specific genome distributions and sample exposures to quantify the selection coefficients (neutral, positive, or negative) acting on functional up-regulation or inactivation for different genes. This approach is limited to simple SV classes—

such as translocation, foldback, reciprocal inversion, as well as deletion and tandem duplication shown in Figure 6.1—subject to appropriate estimation of the tissue-specific rearrangement rate at every position in the class-specific 2D annotation map.

To estimate the SV probability at every point (or pixelated square for reduced computation) in the 2D map (triangle for intra-chromosomal events; rectangle for inter-chromosomal events), recall that $\Pr(A \cap B) = \Pr(A) \Pr(B \mid A)$. In this context, the probability of a BPJ between positions (or pixels) A and B is the marginal breakpoint probability at A, multiplied by the conditional probability of a partner break at B. The first factor is easily obtained via class-specific logistic regression models explored in Section 3.3. The second factor is harder to obtain, and depends on the class (or signature) size distribution, sequence homology, physical proximity imposed by TAD structure and neighbouring chromosome territories, and the marginal breakpoint likelihood of B for this SV class. If this proves intractable to estimate, another possibility is to eschew 1D breakpoint likelihood models (such as logistic regression) in favour of 2D spatial point process models for the event locations directly observed within the space of possible junctions. For the spatial point process, the predictor variables at each 2D location could include size, homology, proximity (from Hi-C data), and a range of properties along the 1D genome that somehow require translation to the 2D junction space. In either scenario, properties such as chromatin state and gene expression should ideally be regarded as tissue-specific predictors. With a spatial point process, it may even be beneficial to regard tissue type as a third dimension, along which some tissue-agnostic properties are held constant, and the tissue-specific properties allowed to vary by identity pixels sorted by relatedness of tissue development and/or chromatin correlation.

One important caveat to using observed cancer variation datasets as the basis for background rearrangement rate models is the disproportionate bias towards positively-selected driver events, as previously discussed in Section 3.6. A more critical limitation is that background rate models do not readily extend to complex structures involving several genome loci in convoluted configurations. If the annotated consequences of templated insertion, chromoplexy, chromothripsis, and other structures, cannot be modelled as probabilistic distributions in the absence of selection, it is difficult to conceive how recurrence-based driver analysis will be possible without massive simplification. In the short to medium term, the prospects for driver discovery with complex SV may be limited to existing approaches on a reduced profile—such as copy number (Mermel et al.,

2011) or junction enrichment (Wala et al., 2017a)—*or* depend on functional assessment of related alterations to transcription, translation, and/or chromatin conformation, where complementary data (such as RNA-seq) are available to elaborate on the SV effect. Given the many challenges in interpreting SV structures and consequences, experimental validation of putative drivers is especially pertinent, perhaps using CRISPR technology to recreate specific rearrangement structures with a predicted functional consequence (Maddalo et al., 2014).

6.4.3 Clinical translation

Method development for somatic SV specification and annotation has important clinical ramifications (Macintyre et al., 2016b), with driver alterations *and* signatures of underlying repair deficiency illuminating diagnosis, prognosis, therapeutic opportunities, and the dynamics of ongoing genome instability which facilitates adaptation and acquired drug resistance.

6.5 Concluding remarks

Through errors of DNA repair, replication, and segregation, somatic genomes gradually diverge from their common ancestor in the zygote, occasionally evolving into cancerous cell populations with unregulated growth. Genome alterations at any scale may contribute to oncogenic transformation, with this thesis focussing on structural variation (typically larger than 1 kb) detected through whole genome sequencing of 2559 PCAWG samples. In addition to previously recognised SV phenomena involving isolated junctions of non-contiguous sequence or, at the other extreme, mass rearrangement under catastrophic stress, the PCAWG dataset revealed a vast intervening continuum of medium-complexity structure with hallmarks of both 'break and ligate' and 'template and replicate' repair modalities. By methodically surveying this panorama of SV structures and properties, the available repertoire of genetic manoeuvres is revealed with unprecedented breadth and resolution across dozens of common cancer types. The tissue and sample specificity of SV form, size, location, and complexity are testament to the many diverse rearrangement mechanisms driving somatic genomes towards pathological cancer phenotypes.