PyClone: Statistical inference of tumour evolution from deep digital sequencing

Andrew Roth[1,2], Gavin Ha[1,2], Samuel Aparicio[2,3], Alexandre Bouchard-Côté[4], Sohrab P. Shah[2,3♣]

[1] *Bioinformatics Graduate Program, University Of British Columbia, Vancouver, Canada*
[2] *Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, BC V5Z 1L3, Canada*
[3] *Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, V6T 2B5, Canada*
[4] *Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*
♣ *To whom correspondence may be addressed: sshah@bccrc.ca.*

We introduce a novel statistical method for inference of tumour evolution in human cancers. PyClone is a non-parametric, hierarchical Bayes clustering method to simultaneously group sets of deeply sequenced somatic mutations into putative clonal clusters, estimate their cellular prevalences and account for allelic imbalances expected to be introduced by segmental copy number changes and normal cell contamination. PyClone can simultaneously analyze multiple tumour samples from the same patient, providing unprecedented opportunity to accurately infer clonal genotypes.

Cancer progresses under Darwinian evolution where (epi)genetic variation alters molecular phenotypes in individual cells with implication for positive and negative clonal expansions. Thus tumours consist of heterogeneous clonal cell populations, related through a phylogeny (Figure S1). Distinct clonal cell populations (which can be defined by their genotypes) act as substrates for selection in the context of tumour micro-environments and/or therapeutic intervention [1,2]. However, identifying clonal genotypes within a tumour remains a formidable and unmet challenge. Defining genotypes is an essential step towards understanding biological properties underlying clonal expansions, and ultimately, clinical trajectories. A logical solution is to cluster mutations into putative clonal genotypes assuming that mutations co-occurring in the same cells should be grouped together. This facilitates identification of specific clones by their shared mutations, inference of ancestral clones implicating mutations driving tumourigenesis, and classification of clones on the basis of expansion or contraction under different selective pressures. Consequently the cellular prevalences of mutations (inferred from bulk populations) defining a clonal genotype will shift together as clones expand or contract. This suggests an opportunity to identify clonal genotypes by grouping mutations according to shared dynamic cellular prevalence patterns.

Emergence of deep digital sequencing (DDS) with NGS devices has enabled measuring and modeling clonal evolution in cancer. Digital sequencing of mutations in tumour cell populations [3–8] provides a first approximation to identify clonal population structure. Furthermore, application to multiple anatomic sites [9–11] and/or serially acquired biopsies [3,12] from individual patients permits measuring clonal prevalence shifts of mutations in local micro-environments and over time intervals as a reflection of positive or negative selection.

Despite progress in sequencing technology, statistical approaches to model clonal evolution from deep digital sequencing of mutations remain underdeveloped, with poorly understood analytical assumptions. Accurate inference is currently limited by four major factors: i) variation in allelic measurements due to technical sources; ii) the influence of total and allele-specific segmental copy number from complex karyotypes [13,14] on allelic measurements (e.g. the mutational genotype); iii) mixtures of unknown numbers of subpopulations, each with unknown prevalence and iv) lack of capacity to simultaneously analyze multiple datasets for identifying clonal genotypes.

To systematically address these deficiencies, we developed PyClone: a novel, hierarchical Bayes statistical model (Figure S1, Figure S2). The inputs to the model are a set of deeply sequenced mutations (optionally from multiple samples) and a measure of allele-specific copy number at each mutation locus (obtained from analysis of whole genome sequencing or high-density genotyping arrays (Figure S3, Supplementary information)). The outputs consist of the posterior densities over the cellular prevalences for each mutation and the probability of the co-clustering of each pair of mutations. The framework overcomes the the limitations outlined above through four novel modelling advances: incorporating over-dispersed emission densities; flexible priors over mutational genotypes (accounting for copy number when estimating cellular prevalence); Bayesian non-parameteric clustering to discover groupings of mutations; and joint inference over multiple samples to derive clonal genotypes. The fully featured model along with simpler derivatives used for benchmarking comparisons are fully detailed in the Supplementary information. Simulated data sets systematically illustrate improvements in performance (Figure S4, Supplementary information). In particular, use of the parental copy number (PCN) prior consistently gave the best performance for both mutational clustering and cellular prevalence estimation.

We evaluated our approach on idealized DNA mixtures (Figure 1a), with datasets produced by mixing material from four 1000 genomes project [15,16] individuals (Supplementary information). Each mixture contained DNA in approximate proportions of 0.01, 0.05, 0.20, 0.74. Specific single nucleotide variants were amplified using PCR, and then sequenced deeply on the Illumina MiSeq platform. We could therefore establish ground truth (Figure 1b) by selecting positions with variants found in only one of the cases. PyClone using the PCN prior significantly outperformed all other methods (Figure 1a). Modelling over-dispersion of the count data significantly improved performance for both PyClone and genotype naive clustering. Joint analysis with the PCN Beta-Binomial model achieved perfect performance (V-measure=1.0) in all 10 sub-sampled datasets (Figure 1a). Accounting for mutational genotype and joint inference each conferred independent performance gains. To illustrate the effect of mutational genotype, we randomly selected one of the ten runs and present the detailed results contrasting PyClone with an infinite Beta-Binomial mixture model (IBBMM, Supplementary information), a baseline analogous to clustering the raw allelic data without considering the influence of mutational genotype. The IBBMM model output

8 clusters corresponding to the 4 sets of heterozygous (genotype *AB*) and 4 sets of homozygous (genotype *BB*) mutations (Figure 1c). By contrast, PyClone identifies the 4 correct clusters (Figure 1d), placing both the AB and BB mutations from the same clones together (Figure 1e).

Having established improvements in accuracy on idealized data, we then assessed the multi-sample PCN and IBBMM models in the cancer setting, using recently published mutational profiles of multiple samples from two high grade serous ovarian cancers (HGSOC) [10]. (Note, to illustrate the utility of the single sample PCN model, analysis of two triple negative breast cancers are shown in the Supplementary information, Figure S5, Table S1, S2.) Four spatially separated samples taken from each of two primary ovarian tumours (Case1 and Case2) were evaluated. Mutations called in exomes were deeply sequenced (~5000x) with targeted amplicon sequencing, yielding sets of 72 validated somatic mutations in Case1 and 49 mutations in Case2. Copy number priors were obtained from high-density genotyping arrays (Supplementary information). We expected the cellular prevalences of mutations present in the same clone to vary together over the spatial samples. By corollary, covariation of cellular prevalences would likely reflect mutations sharing the same clonal genotypes.

Case1 mutations clustered by IBBMM (Figure 2a,b, Table S3) resulted in mutations with the highest allelic prevalences (Cluster 1, $n = 9$ mutations) grouped together. These mutations showed a similar prevalence pattern to mutations in Cluster 5 ($n = 31$ mutations) across samples. We suggest these mutations belong to the same clone, with Cluster 1 mutations predominantly homozygous and Cluster 5 mutations predominantly heterozygous (Table S3). Eight of nine mutations in IBBMM Cluster 1 fall into regions of heterozygous deletion in all four samples. By contrast, 28 of 31 mutations in Cluster 5 are in diploid heterozygous regions (Table S3) in all four samples. Therefore, the difference in cellular prevalence estimates in IBBMM between Cluster 1 and 5 can be explained by the copy number of the loci spanning the mutations impacting the mutational genotype. PyClone instead groups the mutations corresponding to IBBMM Cluster 1 and Cluster 5 into one group of $n = 44$ mutations (Figure 2c,d - Cluster 1), with representation of both the heterozygous and homozygous loci at cellular prevalences of near 1.0. Case2 results showed a similar pattern (Table S4). The IBBMM results (Figure 2e,f) in twice as many clusters as PyClone (Figure 2h,i). Cluster 1 ($n = 4$ mutations) had the highest inferred cellular prevalence, but with a similar pattern across the samples to mutations in Cluster 2 ($n = 20$ mutations). Three of four mutations in Cluster 1 were in homozygous regions whereas 18 of 20 mutations in Cluster 2 were in diploid heterozygous mutations. PyClone clustered these 24 mutations plus one additional mutation into a single group (Cluster 1, $n = 25$ mutations) with cellular prevalence near 1.0.

PyClone Cluster 1 in both Case 1 ($n = 44$ mutations) and Case 2 ($n = 25$ mutations) (Figures 2d,h) likely represent mutations comprising the ancestral clone in these tumours' aetiology. In contrast, clusters with cellular prevalences lower than 1.0 indicate putative descendant clones. In Case 2, PyClone predicted four clusters. We suggest the following interpretation: Cluster 1 represents the ancestral clone, while Clusters 2, 3, 4 represent descendent clones, which expanded in regions D, B and C of the tumour respectively. We emphasize that IBBMM splits the ancestral clone into at least two clusters in both Case 1 and Case 2, with evidence from the copy number analysis (Table S3, S4) attributing the split based on heterozygous or homozygous mutational genotype. PyClone modelling of mutational genotypes coupled with simultaneous inference across multiple samples therefore likely provides a more robust approach to ascertaining clonal genotypes with dramatic implications for how cellular prevalence estimates of mutations are interpreted in reconstruction of evolutionary histories.

In summary, we have introduced a novel statistical approach for improved inference of clonal evolution patterns in human cancers from deep digital sequencing of mutations. Discussion on the limitations, future directions and generalizability of the approach are included in the Supplementary information. The advances we present have practical implications for inference of clonal genotypes and show measurable reductions in spurious inference relative to current approaches. As the practice of measuring allelic prevalences during clinical course [17,18] or through retrospective analysis of multiple samplings increases [9–11], we suggest that PyClone will contribute a robust statistical inference approach for studying selection patterns underpinning disease progression in cancer.

Methods

Methods are detailed in the Supplementary information.

Acknowledgments

This work is funded by the Canadian Institutes for Health Research (CIHR), Genome Canada and the Canadian Cancer Society Research Institute. SPS is supported by the Michael Smith Foundation for Health Research. AR is supported by a CIHR Banting scholarship.

References

1. Greaves, M. and Maley, C.C. *Nature* **481**, 306–313 (2012).

2. Aparicio, S. and Caldas, C. *The New England journal of medicine* **368**, 842–851 (2013).

3. Shah, S.P. *et al. Nature* **461**, 809–13 (2009).

4. Ding, L. *et al. Nature* **464**, 999–1005 (2010).

5. Govindan, R. *et al. Cell* **150**, 1121–34 (2012).

6. Nik-Zainal, S. *et al. Cell* **149**, 994–1007 (2012).

7. Shah, S.P. *et al. Nature* **486**, 395–9 (2012).

8. Carter, S.L. *et al. Nat Biotechnol* **30**, 413–21 (2012).

9. Gerlinger, M. *et al. N Engl J Med* **366**, 883–92 (2012).

10. Bashashati, A. *et al. Journal of Pathology* (2013 (accepted)).

11. Sottoriva, A. *et al. Proceedings of the National Academy of Sciences of the United States of America* **110**, 4009–4014 (2013).

12. Ding, L. *et al. Nature* **481**, 506–10 (2012).

13. Curtis, C. *et al. Nature* **486**, 346–52 (2012).

14. Bignell, G.R. *et al. Nature* **463**, 893–898 (2010).

15. 1000 Genomes Project Consortium *Nature* **467**, 1061–73 (2010).

16. Harismendy, O. *et al. Genome Biol* **12**, R124 (2011).

17. Forshew, T. *et al. Science translational medicine* **4**, 136ra68 (2012).

18. Dawson, S.J. *et al. The New England journal of medicine* **368**, 1199–1209 (2013).

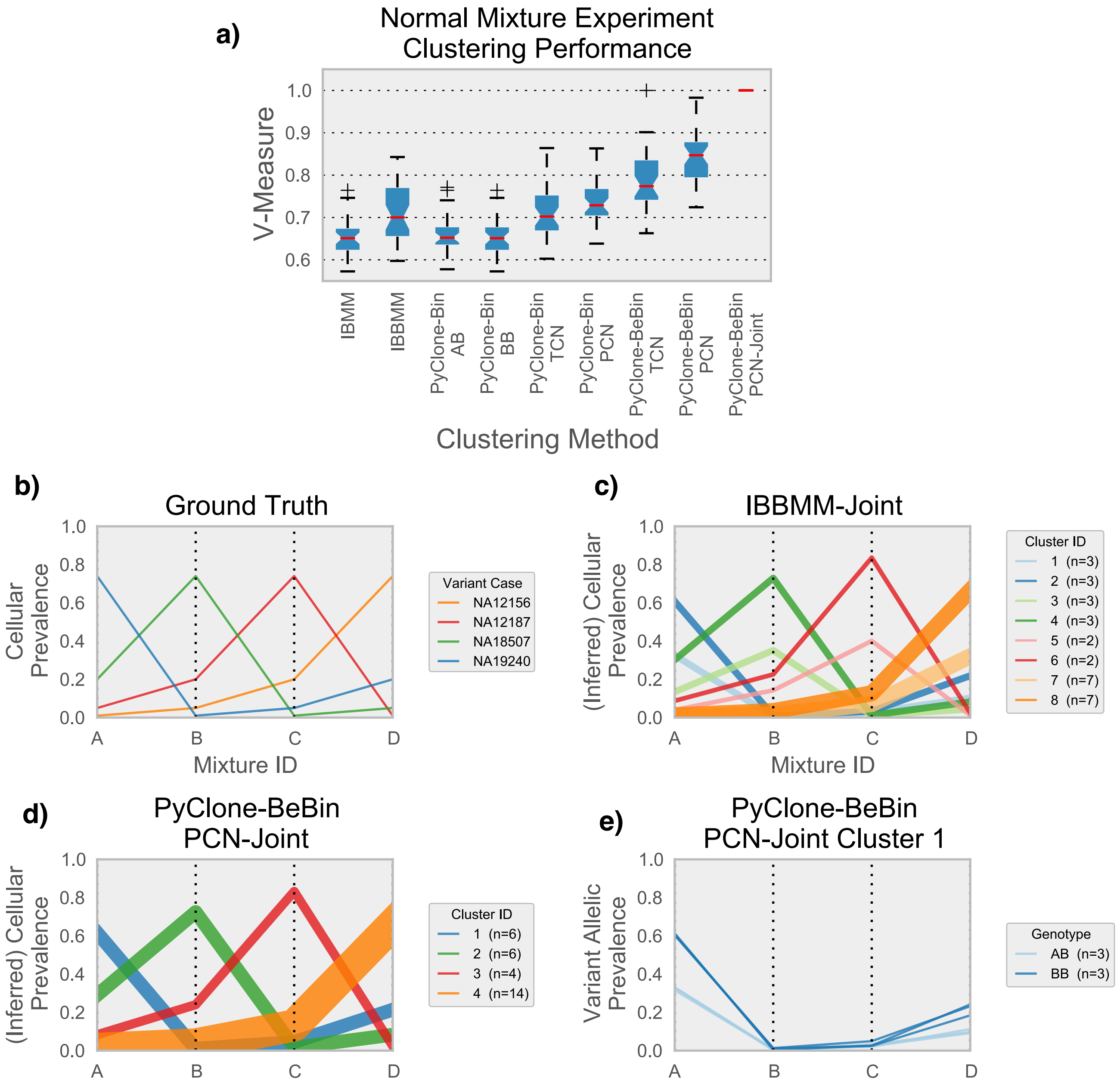
Figures and Tables

Figure 1: **a)** Clustering performance of different methods when applied to the mixture of normal tissues datasets. We created 10 subsets of 30 mutations which were a 50:50 mixture of mutations which had the AB or BB genotype in the variant case. For all methods, except PyClone-BeBin-PCN-Joint, we analysed these 10 subsets of mutations in each of the four mixtures (n=40 mutations). The PyClone-BeBin-PCN-Joint method jointly analysed each subset of mutation across all four mixtures (n=10). **b)** Expected cellular prevalence based on mixing proportions for the normal tissue mixing experiment. Clusters are color coded by the case which was predicted to posses the variant genotype (either AB or BB). Inferred cellular prevalence for each of the clusters identified when using the **c)** IBBMM and **d)** PyClone methods to jointly analyse all four samples (mixtures). The cellular prevalence of the cluster is the mean value of the cellular prevalence of mutations in the cluster and the width of the lines for each cluster is proportional to the number of mutations in the cluster. The number of mutations *n* in each cluster is shown in the legend in parentheses. **e)** variant allelic prevalence for each mutation in cluster 1 from the PyClone analysis. Mutations are color coded by the genotype of the mutation in the variant case (NA19240). Note that cluster 1 predicted by PyClone consists of mutations with both the AB and BB genotypes which were placed in separate clusters (1 and 2) by the IBBMM.

{fig:mixing}

Figure 2: Joint analysis of multiple samples from high grade serous ovarian cancer (HGSOC) cases 1 and 2. Panels a) and e) show the variant allelic prevalence for each mutation color coded by predicted cluster using the IBBMM to jointly analyse the four samples from case 1 and 2 respectively. Panels c) and g) show the same thing but with clusters predicted by PyClone using the PCN method for setting priors and a Beta Binomial emission density. Panels b) and f) show the inferred cellular prevalence for each cluster using the IBBMM method for cases 1 and 2 respectively. Panels d) and h) show the same thing but using PyClone as discussed above. As in figure 1 the cellular prevalence of the cluster is the mean value of the cellular prevalence of mutations in the cluster and the width of the lines for each cluster is proportional to the number of mutations in the cluster. The number of mutations *n* in each cluster is shown in the legend in parentheses.

{fig:ith}



a)
Normal Mixture Experiment
Clustering Performance
V-Measure
1.0
0.9
0.8
0.7
0.6
IBMM
IBBMM
PyClone-Bin AB
PyClone-Bin BB
PyClone-Bin TCN
PyClone-Bin PCN
PyClone-BeBin TCN
PyClone-BeBin PCN
PyClone-BeBin PCN-Joint
Clustering Method
b)
Ground Truth
Cellular Prevalence
Variant Case
NA12156
NA12187
NA18507
NA19240
Mixture ID
A
B
C
D
c)
IBBMM-Joint
(Inferred) Cellular Prevalence
Cluster ID
1 (n=3)
2 (n=3)
3 (n=3)
4 (n=3)
5 (n=2)
6 (n=2)
7 (n=7)
8 (n=7)
Mixture ID
d)
PyClone-BeBin PCN-Joint
(Inferred) Cellular Prevalence
Cluster ID
1 (n=6)
2 (n=6)
3 (n=4)
4 (n=14)
Mixture ID
e)
PyClone-BeBin PCN-Joint Cluster 1
Variant Allelic Prevalence
Genotype
AB (n=3)
BB (n=3)
Mixture ID

Figure 1

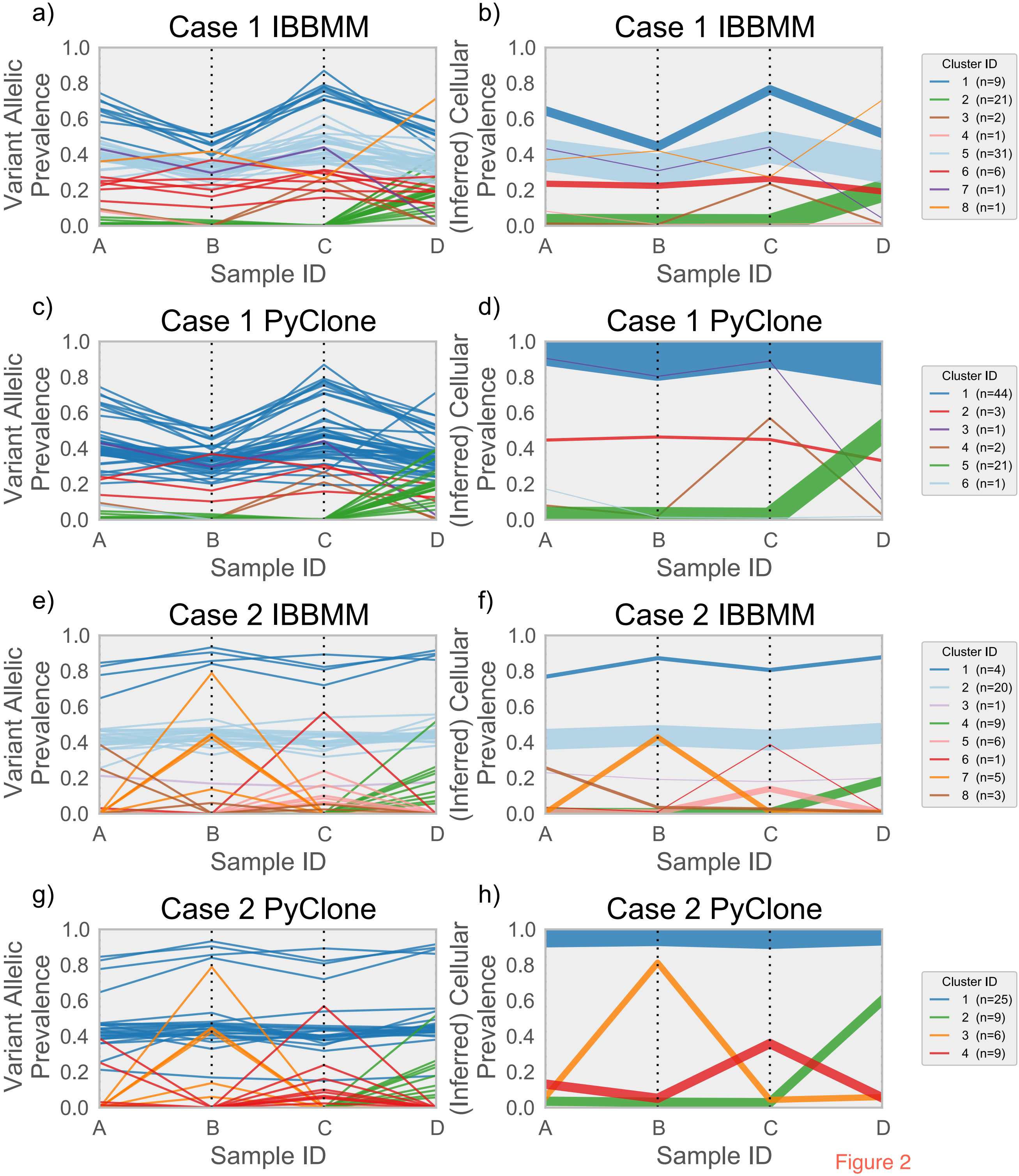

a)
Case 1 IBBMM
Variant Allelic Prevalence
Sample ID
A
B
C
D
1.0
0.8
0.6
0.4
0.2
0.0
b)
Case 1 IBBMM
(Inferred) Cellular Prevalence
Cluster ID
1 (n=9)
2 (n=21)
3 (n=2)
4 (n=1)
5 (n=31)
6 (n=6)
7 (n=1)
8 (n=1)
c)
Case 1 PyClone
d)
Case 1 PyClone
Cluster ID
1 (n=44)
2 (n=3)
3 (n=1)
4 (n=2)
5 (n=21)
6 (n=1)
e)
Case 2 IBBMM
f)
Case 2 IBBMM
Cluster ID
1 (n=4)
2 (n=20)
3 (n=1)
4 (n=9)
5 (n=6)
6 (n=1)
7 (n=5)
8 (n=3)
g)
Case 2 PyClone
h)
Case 2 PyClone
Cluster ID
1 (n=25)
2 (n=9)
3 (n=6)
4 (n=9)

Figure 2