

# Heterogeneity summary, validation and development plans

Brad Chapman

Bioinformatics Core, Harvard Chan School

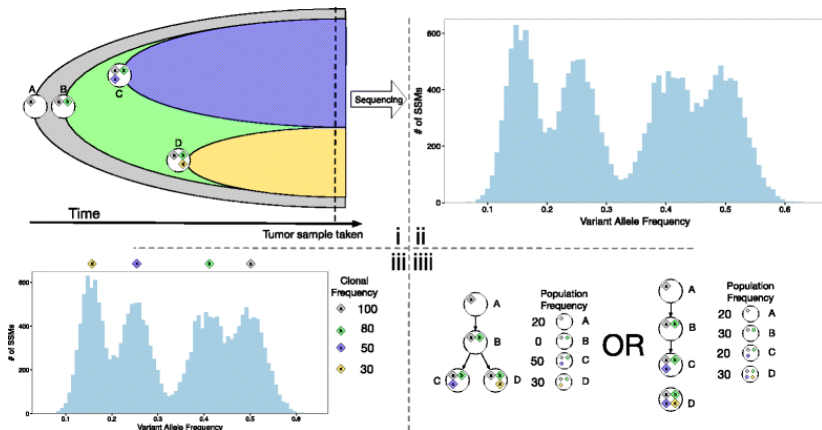
<https://bcb.io>

<http://j.mp/bcbiolinks>

29 March 2017

# Heterogeneity goals

- Automated characterization of tumor purity and clonality
- Support for tumor-only and capture/exome
- Inform SNP, Indel and SV calling



<http://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0602-8>

- Want similar workflow for WGS, capture and tumor-only samples
- Lack of good truth sets
- Tools not fully automated, require manual decision making

# Heterogeneity inputs

- Small variants (SNPs)
- CNVs, either with segmentation (CNVkit) or exons/genes (Seq2C)
- Filter SNP artifacts – UMIs + damage/bias
- Filter CNVs with blacklists

<https://github.com/chapmanb/bcbio-nextgen/issues/963>

- Purity and ploidy  
(PureCN, TitanCNA, BubbleTree, Battenberg)
- CNV Major and minor allele copy numbers  
(PureCN, TitanCNA, Battenberg)
- LOH regions  
(PureCN, TitanCNA)
- Assignment of tumor-only variants to somatic/germline with allele frequencies  
(PureCN)

- Reconstruction: subclones + evolution
  - PhyloWGS (TitanCNA or Battenberg input + variant calls)
- Subclonal identification
  - SciClone (Copy number + variant calls)
  - Guan UofM SMC-Het winning algorithm

## Planned implementation

- PureCN – handles tumor-only and capture with process matched normals; provides purity/ploidy, LOH
- TitanCNA – WGS/exome; purity/ploidy + LOH + allelic CNVs
- PhyloWGS – take TitanCNA input and produce clones and phylogenies
- BubbleTree – supplementary: purity/ploidy + clonal analysis



- BubbleTree – integrated with CNVkit inputs
- PhyloWGS – integrated with Battenberg inputs (WGS only)
- Initial validation work done with PureCN compared to ABSOLUTE and BubbleTree

# Outputs are complex

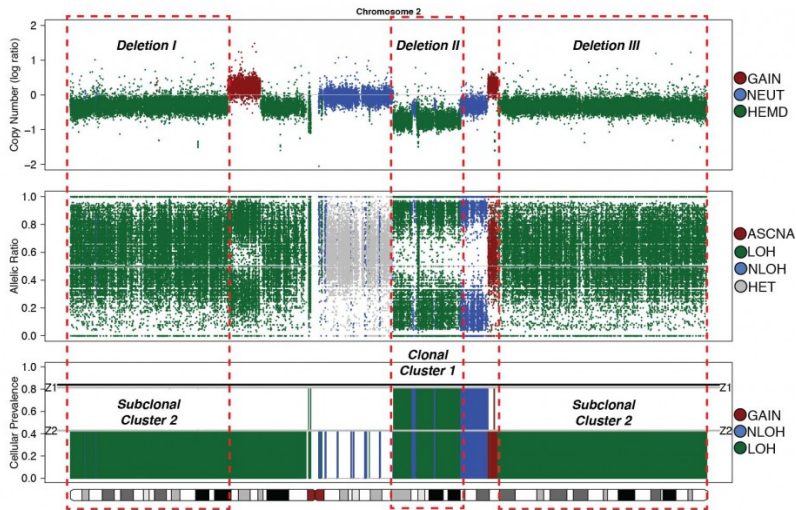
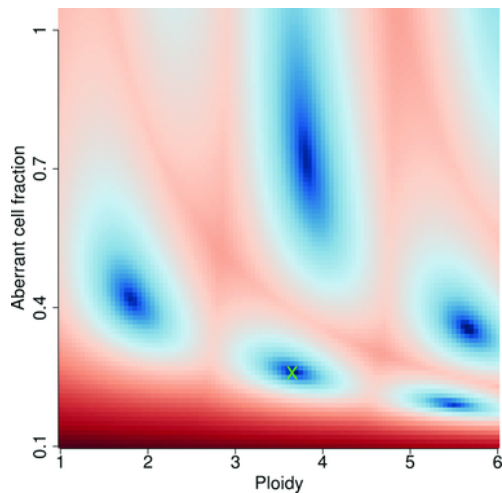


Figure 1

<http://genome.cshlp.org/content/24/11/1881.full>

# Multiple potential solutions



<http://cda.currentprotocols.com/WileyCDA/CPUnit/refId-bi1509.html>

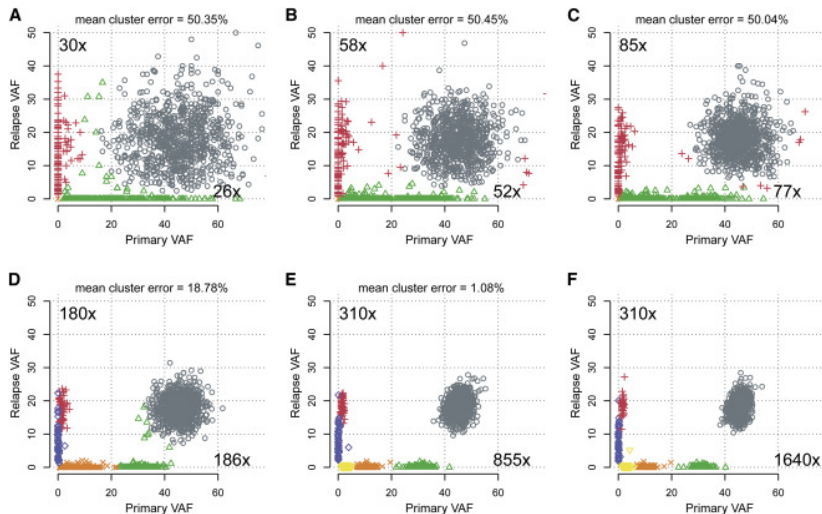
- Currently have good validations for small variants/indels
- CNVs and SVs less established, more difficult
- Genome in a Bottle NA24385 crowdsourced CNVs, subset to exome regions (60 deletions)  
<http://biorxiv.org/content/early/2016/12/13/093526>
- HCC2218 breast carcinoma cell line/blood exomes (43 deletions, 67 duplications)  
<https://github.com/Illumina/Canvas#demo-tumor-normal-enrichment-data>

# Heterogeneity validation

- tHapMix – somatic genome simulator
- purity
- multiple clones
- evolutionary history of clones

<https://github.com/Illumina/tHapMix>

# Sequence deeply enough



# Process matched normal BAMs

- Critical for tumor-only samples
- CNVs: controls for  $\log_2$  depth ratios
- Establish germline heterozygous SNPs – PureCN can estimate based on purity/ploidy in addition to being in public databases