BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Bioinformatics Series: Designing An NGS Study For My Biological Question

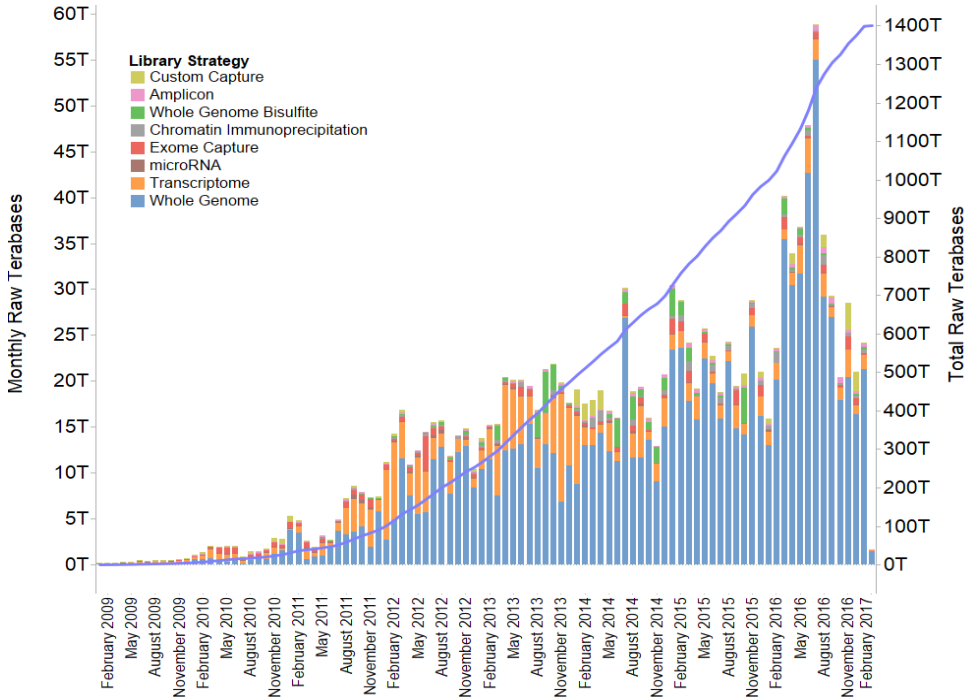
Yussanne Ma
Genome Sciences Centre

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Genome Sciences Centre

Sequencing
5 Illumina HiseqX
4 Illumina HiSeq2500
2 NextSeq500
3 Illumina Miseq
1 Life 3730 xl

1500 libraries per month
>80Tbases per month



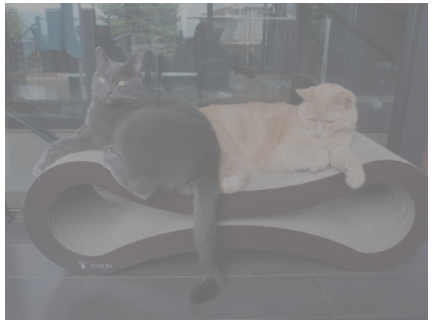
Compute
2 secured data centres
Compute clusters : 800 nodes, 24,000 hyper-threaded cores
48 - 384 GB RAM per node
High memory (1.5TB RAM) computers
>11 Petabytes on-line disk storage

Engaged in over 50 ongoing projects and collaborations from experimental design to data interpretation

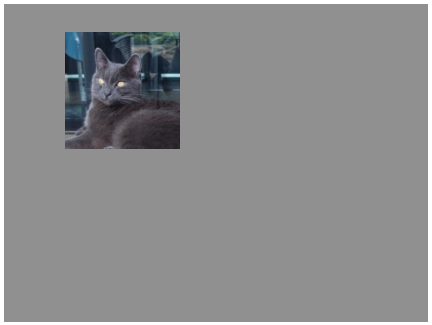
BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Why experimental design is important

Analysis and interpretation of sequencing data is completely dependent on everything upstream



Sample quality, sequencing and type of sequencing matter



The area being sampled matters

Bioinformatic corrections can be made but it's always best to plan ahead

Outline

Genome sequencing

Transcriptome sequencing

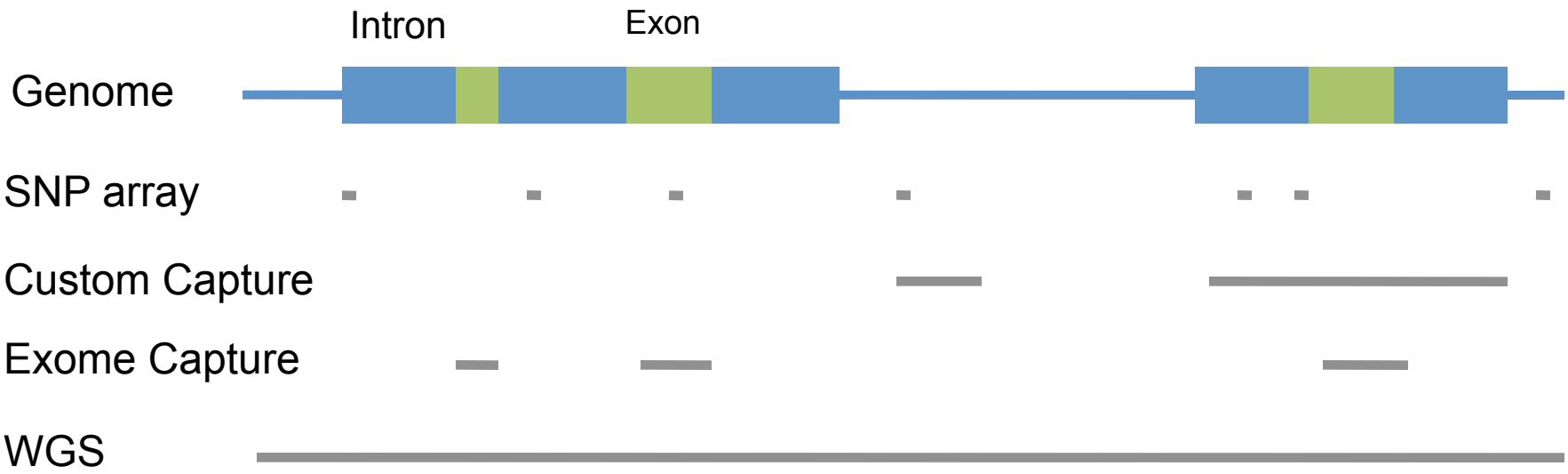
Integrative approaches

Other technologies

Outline

- Genome sequencing
  - Genotyping arrays
  - Exome and custom capture
  - Amplicon
  - Whole genome sequencing
  - Population size and controls
  - Factors affecting quality of variant calls

Genome sequencing overview

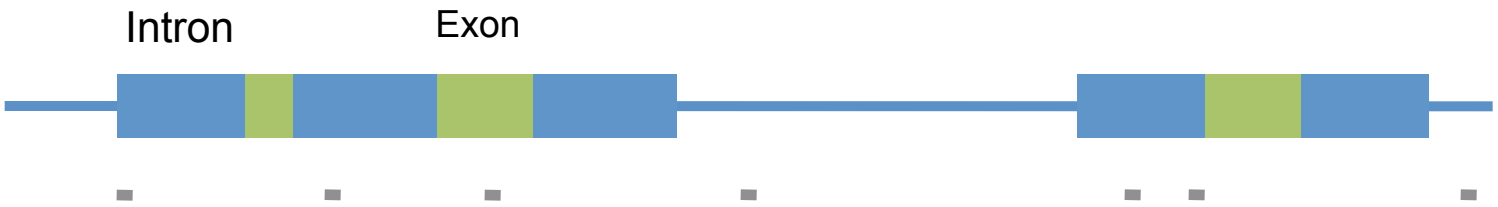


There are many ways to subsample the genome
The cost trade-off is between area covered and depth

Sometimes the genome can be overkill

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Genotyping arrays



Sampling of the genome at locations of known single nucleotide polymorphisms using intensity probes

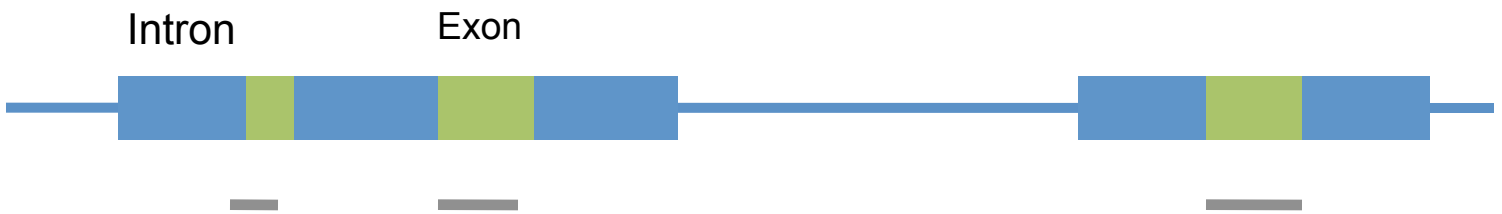
Used for: Studying common variants in large number of cases and controls

Limitations: Cannot be used for calling of rare or novel variants, or structural variants. Resolutions for copy number variation are low

Example project: Genome-wide association study to look for inherited cancer susceptibility loci

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Exome and custom capture



Probes are used to capture all exons, or a specific set of genomic regions

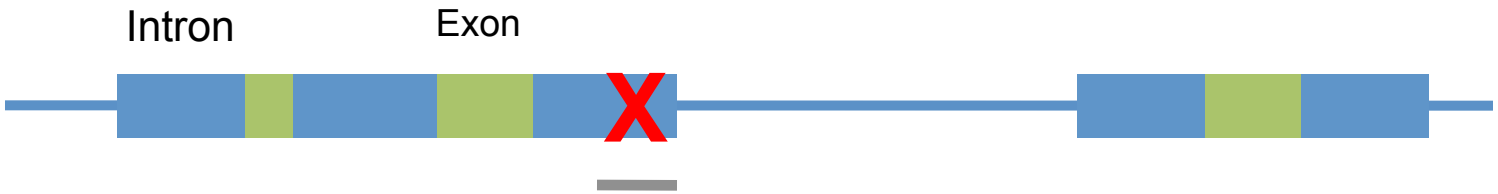
Used for: Studying only coding changes, or a those found in the pre-defined area of interest

Limitations: Cannot call variants outside of capture area. Copy number and structural variants are difficult to call

Example project: Discovery of recurrently mutated genes in large cohort, clinical panel

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Amplicon and Sanger sequencing



Primers are designed that span a genomic event, or sequence across an event.

Used for: Determining the presence or absence of specific events (SNVs, indels, SVs)

Limitations: Cannot be used for discovery, need exact breakpoints in most cases

Example project: Orthogonal verification of putative event discovered by WGS to benchmark tools, determining presence of metastatic fusion event in primary sample.

Whole genome sequencing

Used for: Full characterization of genome including

Novel genes and events

- SNVs and indels not seen in the population
- Private events in recurrent genes or pathways

Complex events

- Copy number
- Structural variants

Genomic landscape of a population

- Mutation signatures

Limitations: Sample size and depth due to cost

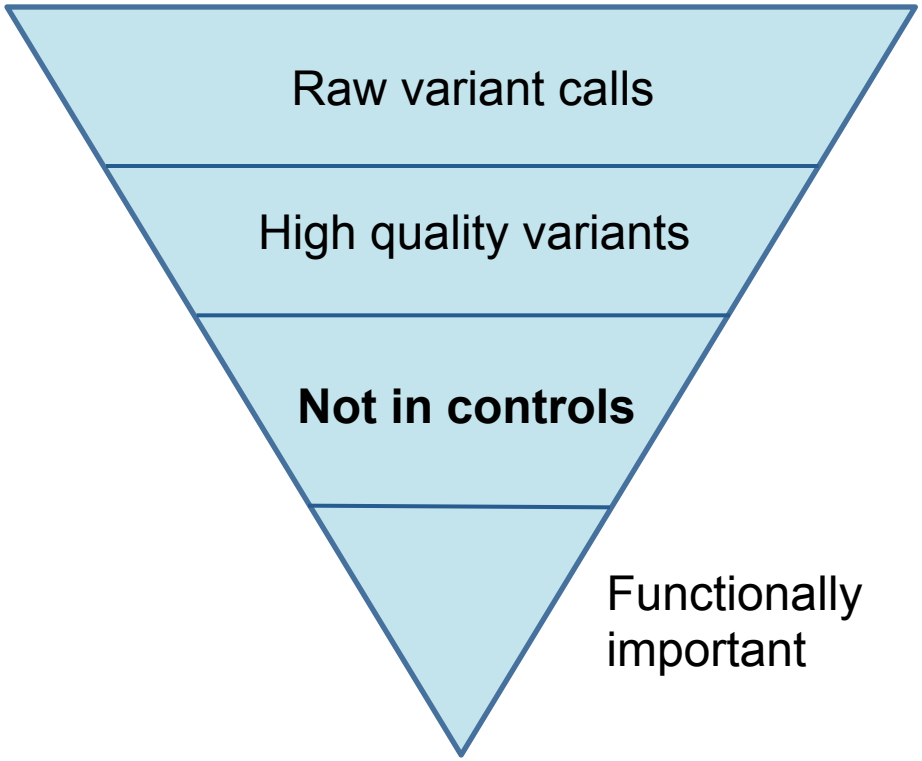
Best used for studies with no *a priori* knowledge of population samples, in depth study of single patient tumour

Genome sequencing: summary

Technology	SNVs	CNVs	SVs	Mutational burden	Mutational landscape
WGS	+++	+++	+++	+++	+++
Exome	+++ (coding only)	+ (coding)	+ (coding)	++	-
Custom capture	+++ (on target)	+ (on target)	+ (on target)	+	-
Genotyping array	Specific events only	+	-	-	-
Amplicon	Specific events only	-	Specific events only	-	-
Sanger	Specific events only	-	Specific events only	-	-

Population size and controls

3 million germline variants, 10,000-100,000 somatic variants on average per sample



Population size and controls

GWAS: Large sample size needed to achieve statistical significance, 1:1 cases and controls

Rare disease: Sequencing of parents reveals patterns of inheritance, sequencing of unaffected relatives helps to filter out passengers

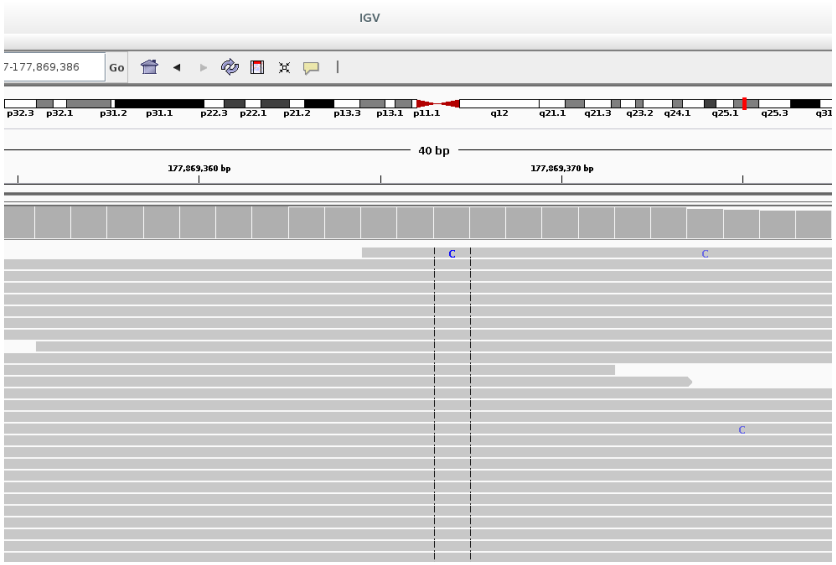
Somatic variants: Matched normal is needed to filter out passenger mutations

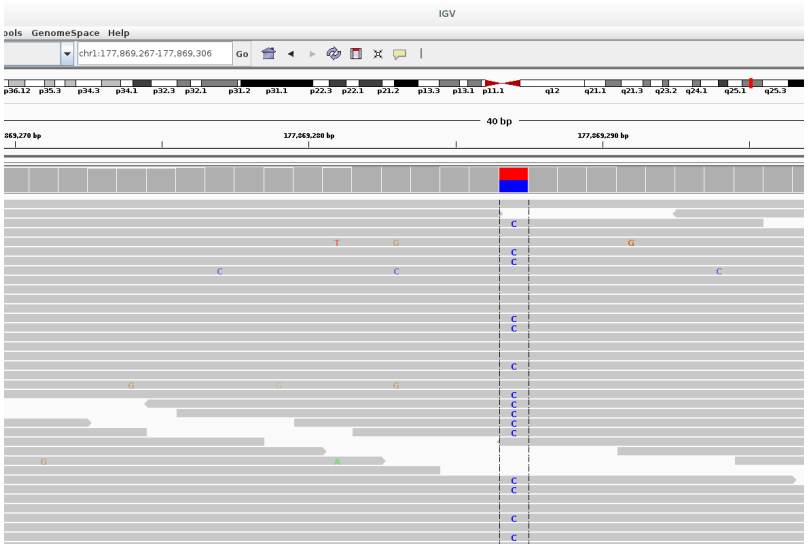
Factors affecting quality of variant calls: sequencing depth

Why 30X genome?

Rule of thumb: it takes 3-10 high quality reads to call a variant

Need to account for variable coverage, evenness of coverage, tumour content, ploidy





Factors affecting quality of variant calls: sequencing depth

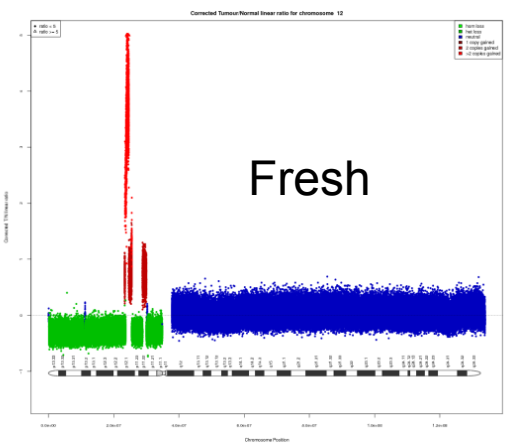
Type of variant	Depth needed
Germline, diploid	30X
Tumour > 70% tumour content	30-40X
Tumour 40 - 70%	40-60X
Low tumour content, subclonal events	> 100X

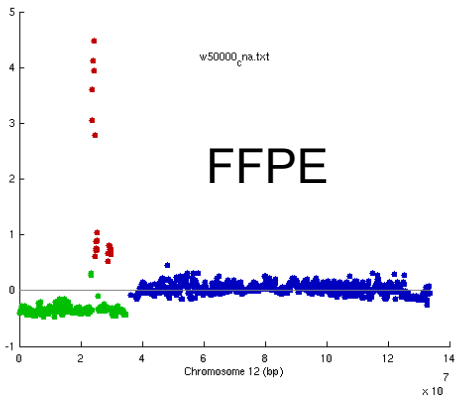
BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

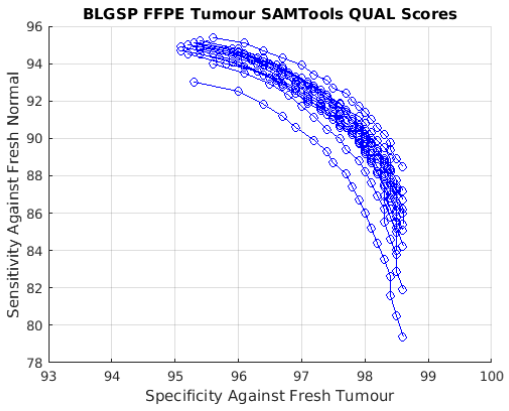
Factors affecting quality of variant calls: sample type

FFPE vs Fresh frozen

All of our protocols (WGS, RNA, miRNA) can be run on FFPE samples, but they may result in slightly lower yield and diversity, and a higher false positive rate for SNV and SV detection, as well as noisier CNV calls







Outline

- Transcriptome sequencing
  - RNA sequencing
  - miRNA sequencing
  - Batch effects

RNA sequencing

Ribosomal depletion vs. polyA selection

no ribosomal RNA captured	higher ribosomal RNA content
non-polyadenylated transcripts are captured	only polyadenylated transcripts are captured
lower minimum input requirement	higher minimum input requirement
higher intergenic and intronic content	lower intergenic and intronic content

RNA sequencing

Used for:

Gene, exon and isoform-level quantification

Quantifying expression of genomic events (SNVs, SVs)

Detecting novel transcripts

Detecting RNA edits

Differential expression between groups (condition/tissue/ tumour type) to identify expression markers

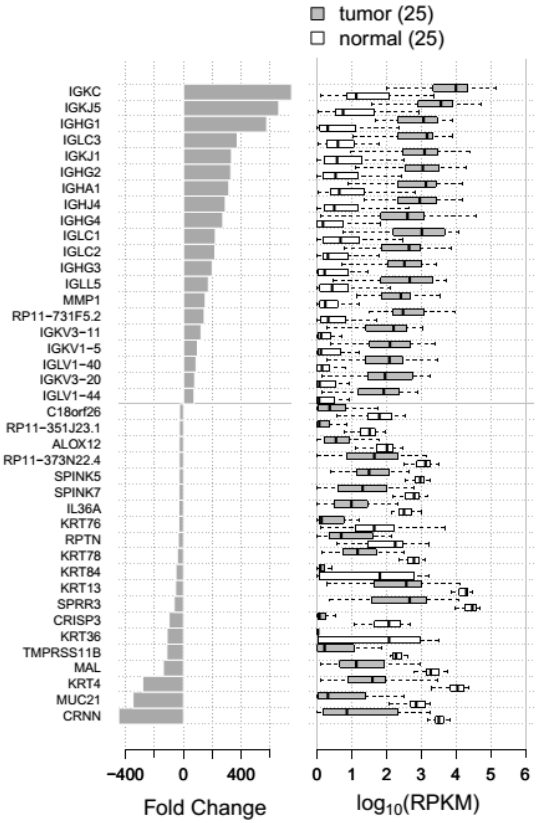
Correlation and clustering of samples by gene expression to identify subgroups

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

RNA sequencing

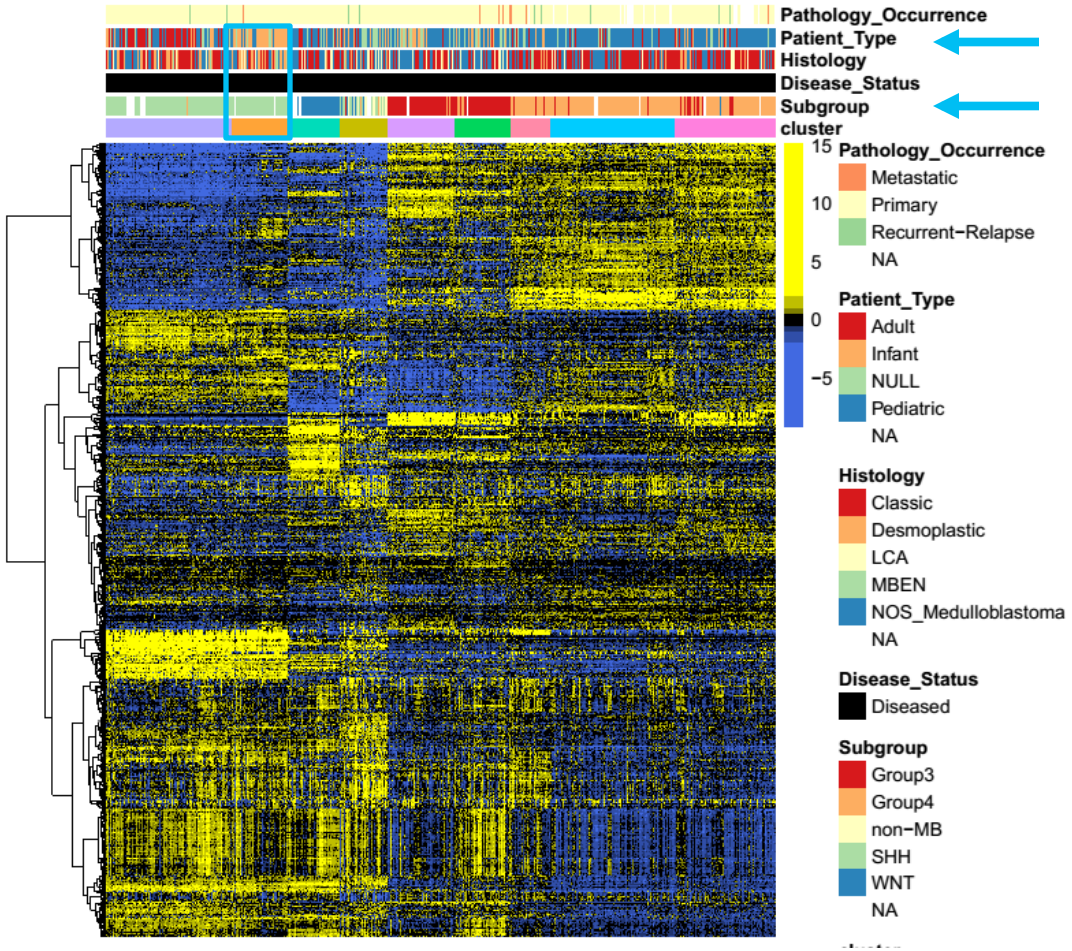
Differential expression between tumour groups

Results are more difficult to interpret with low sample size



BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

RNA sequencing



Hierarchical clustering of medulloblastoma samples by gene expression

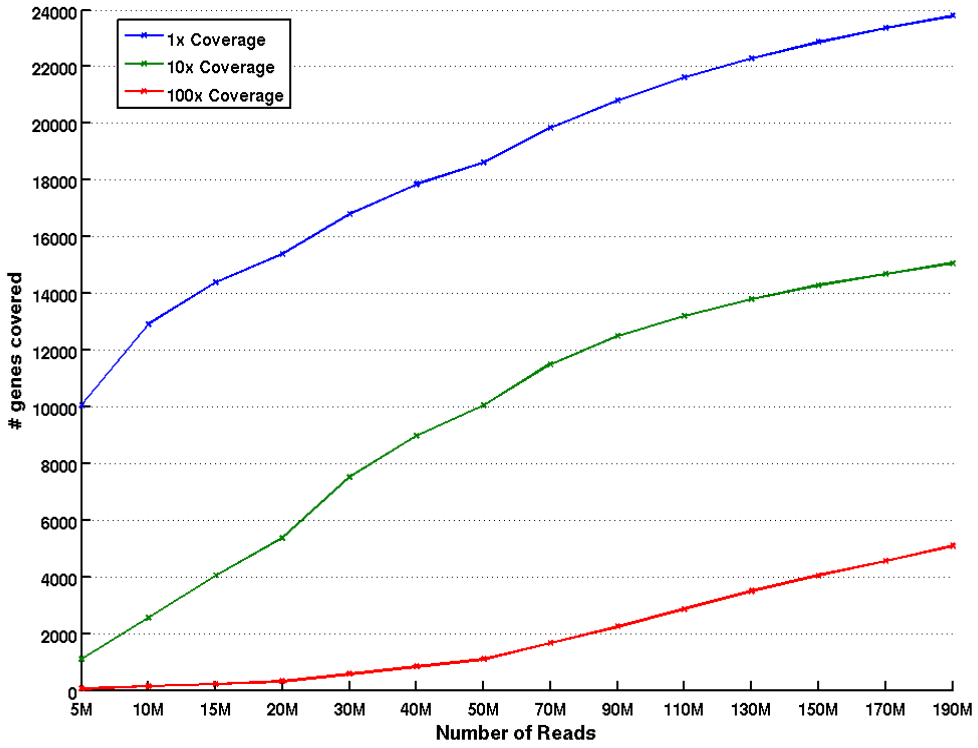
Samples cluster by subgroup

Within a subgroup samples cluster by patient type

Most informative with large sample size and detailed covariates eg. clinical data

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Sequencing depth



Gene diversity for UHR control at different levels of downsampling

Diversity does not tend to saturate

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Sequencing depth

Number of reads per library	200M	120M	60M	40M
Number of genes at 1X	23,000	20,000	18,000	15,000
Number of genes at 10X	14,000	12,000	10,000	5,000
Expression quantification	++	++	++	++
Differential expression	++	++	++	++
Known transcript quantification	++	++	++	++
Detection of structural variants with gene partners or breakpoints specified	++	++	++	+
Detection of SNVs and small indels with known coordinates	++	++	++	++
De novo SNV calling	++	++	+	-
De novo structural variant calling	++	+	Alignment based only	-

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

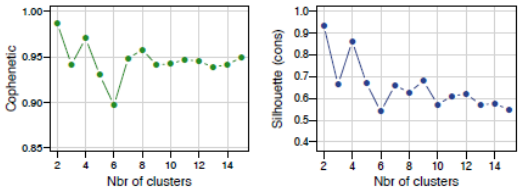
miRNA sequencing

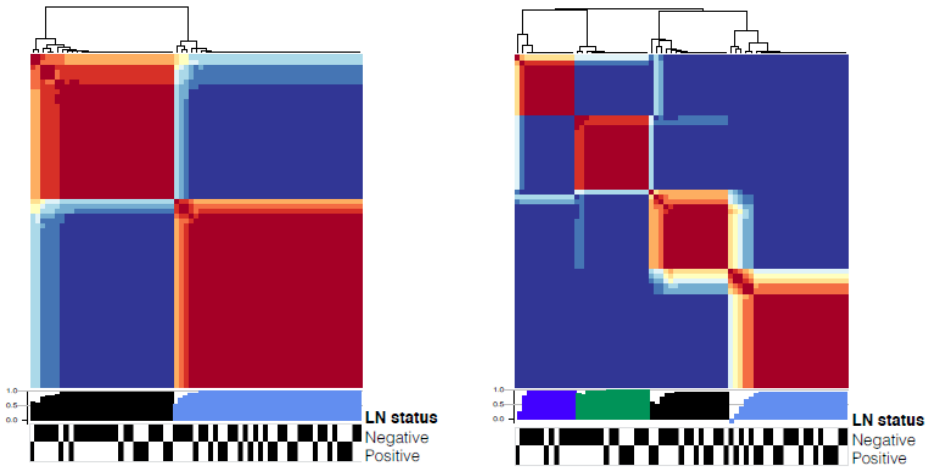
Used for:

Quantification of miRNA expression

Differential expression and expression clustering

Correlation with gene expression to identify targets





Using miRNA to determine a signature for prognosis

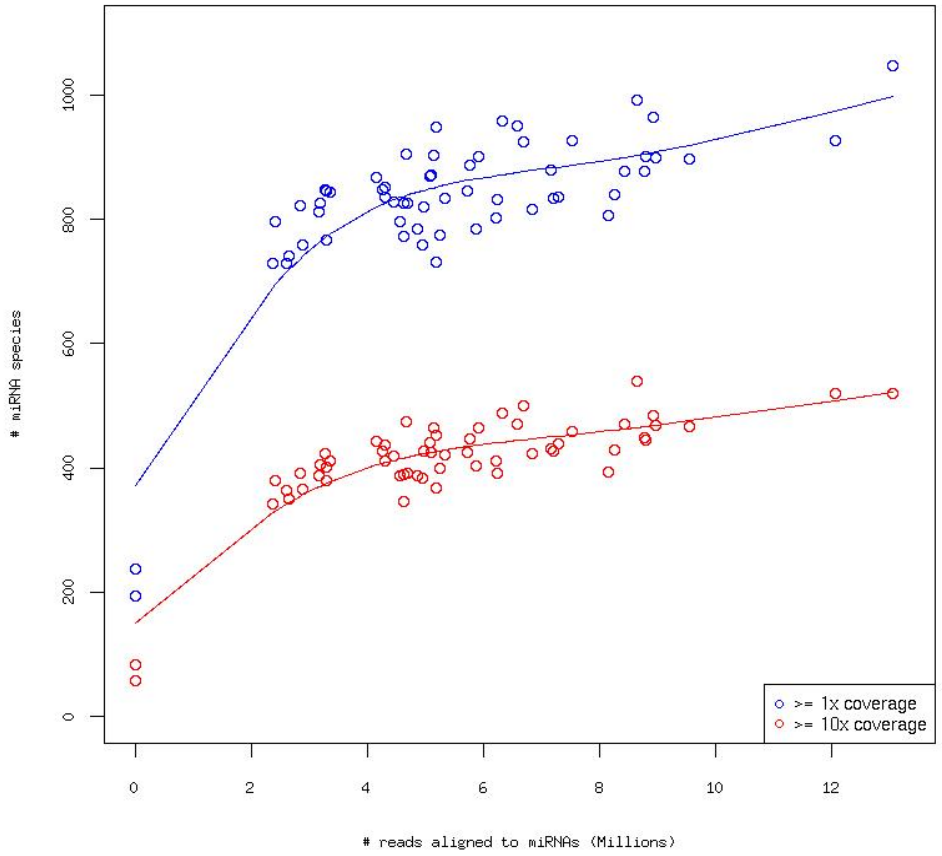
miRNA clustering is found to be more sensitive to subgroups

Search space is smaller and cost of sequencing is lower

May be easier to translate into clinical test

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Sequencing depth



miRNA diversity vs number of reads aligned to miRNA

Two failed samples on far left

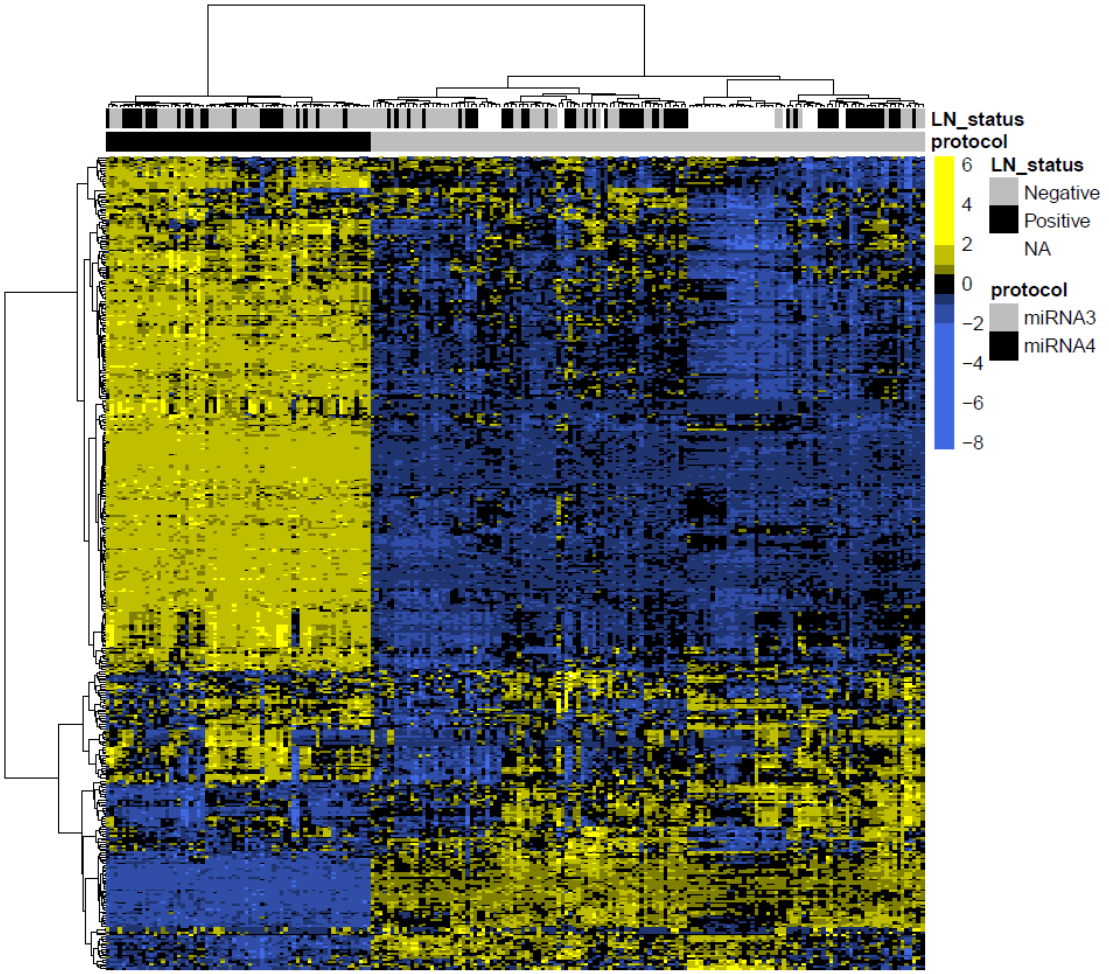
Saturation between 2 and 4 million reads

Batch effects

Samples cluster by protocol, so clustering is difficult to do across multiple protocols

Sample sets sequenced using different protocols are best used as validation, or for meta analysis

Batch effect correction is the most effective with technical replicates



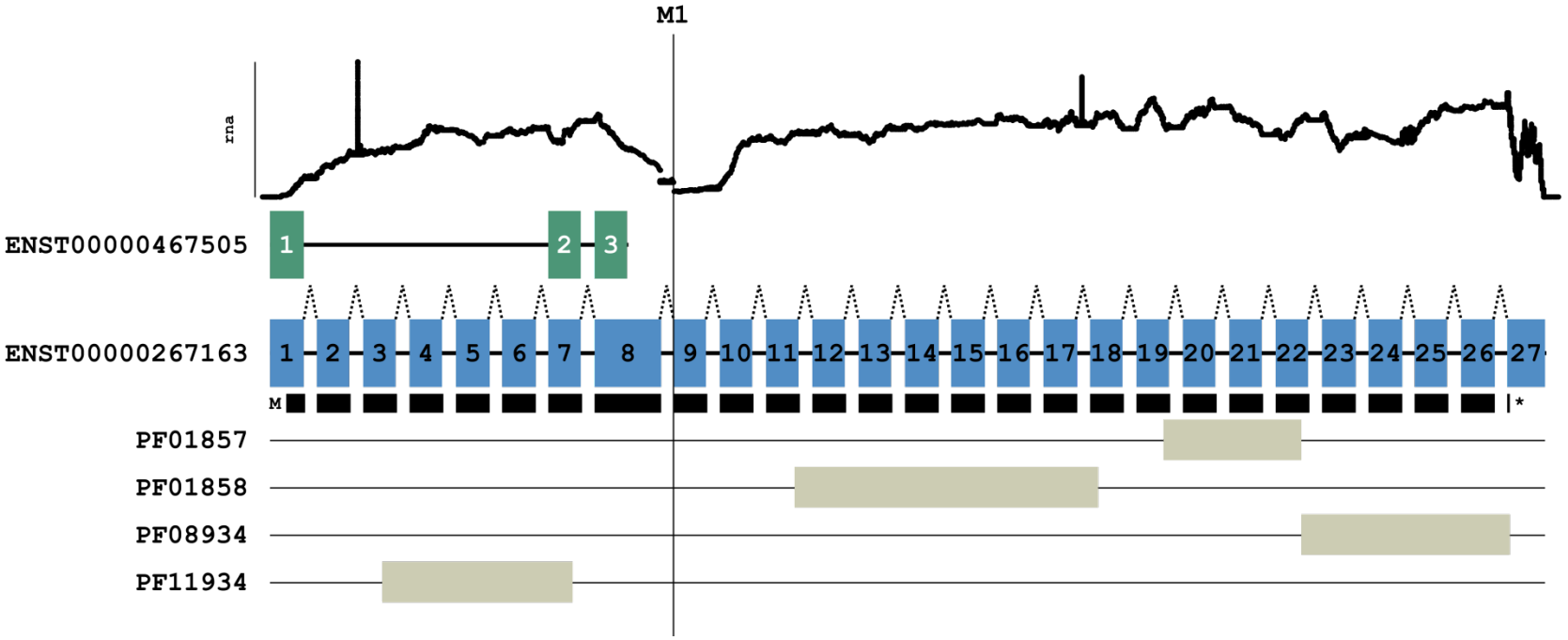
Outline

Integrative approaches

- Genome and transcriptome sequencing
- Clonal evolution experiment
- Integrative analysis to study ‘dark matter’ in cancer

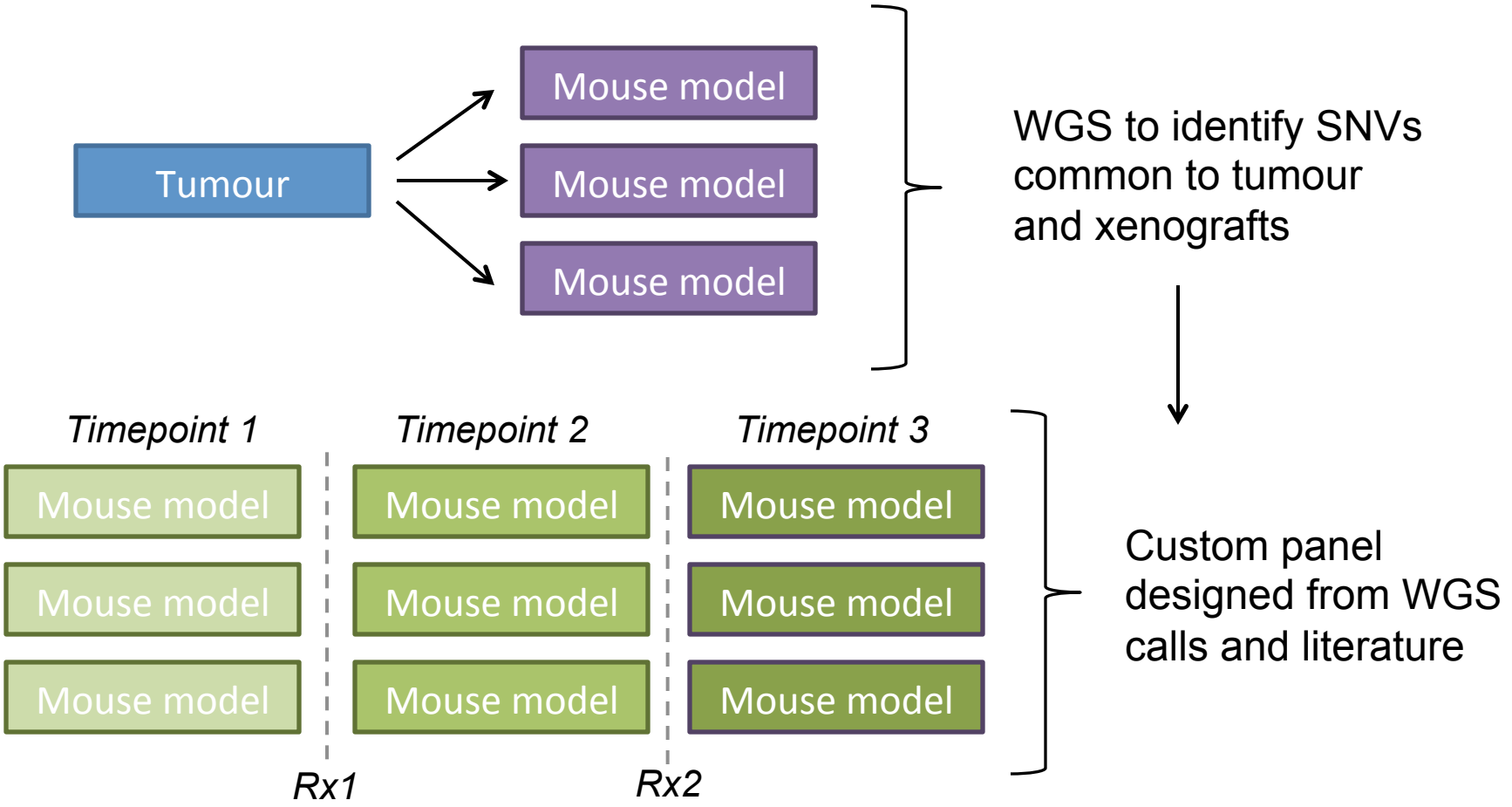
Integrative approaches: genome and transcriptome

RNAseq provides orthogonal validation of genomic events
Combined approach improves specificity, and can identify/confirm alternative splicing and elucidate the effects of genomic events on transcription



Alternative splicing at M1 is identified in the structural variant analysis of RNA and DNA, and gene expression data confirms exon 9 skipping event

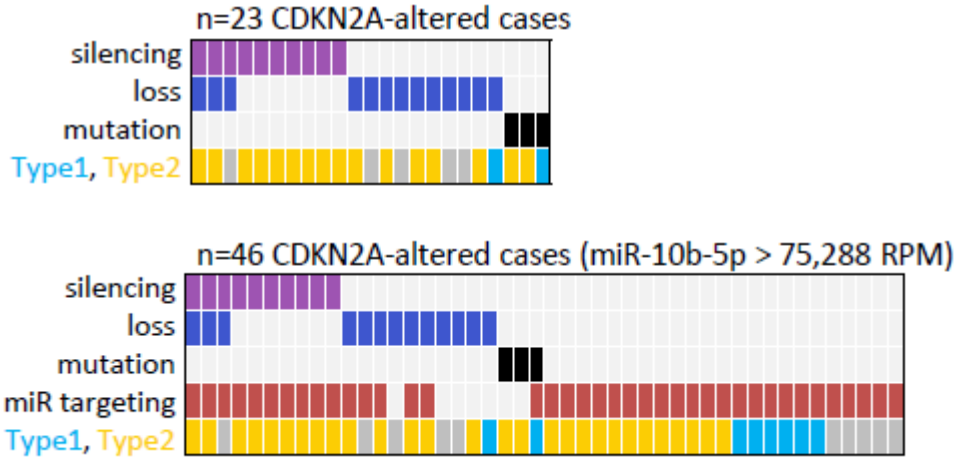
Integrative approaches: Clonal evolution



Clonal evolution over multiple timepoints and treatment events

Integrative approaches: 'dark matter'

miRNA, RNA and WGS and methylation sequencing identify multiple mechanisms in which CDKN2A function is disrupted in papillary renal-cell carcinoma



Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma, The Cancer Genome Atlas Research Network, N Engl J Med 2016; 374:135-145, January 14, 2016

Outline

Other technologies

- Microbial analysis
- de novo genome assembly
- Single-cell sequencing
- Epigenomics
- Immunogenomics

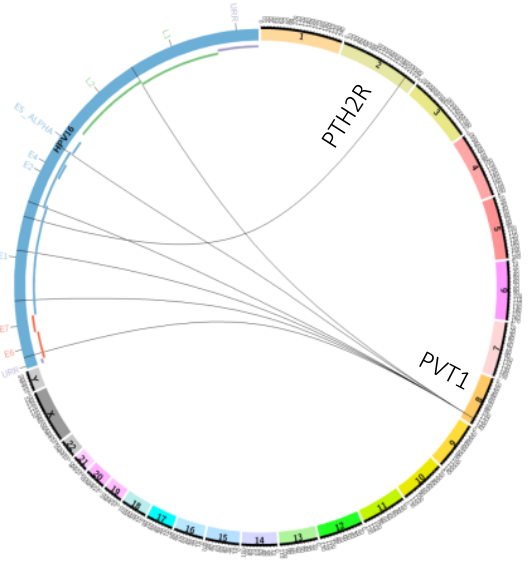
BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Microbial analysis

16S sequencing: Identification and quantification of known bacterial species. Useful for survey of large number of samples

Short read sequencing (shallow): Rapid classification of known microbial species in metagenomic samples

Whole genome and transcriptome sequencing: Microbial expression and genome integration in tumour samples



De novo genome assembly

Short read sequencing at ~30X is sufficient for de novo assembly using ABySS to produce contigs

Contigs can be:

- Aligned to existing references to identify variants in new strains
- Annotated to identify putative genes

Extension to a full draft reference will require additional sequencing to build scaffolds:

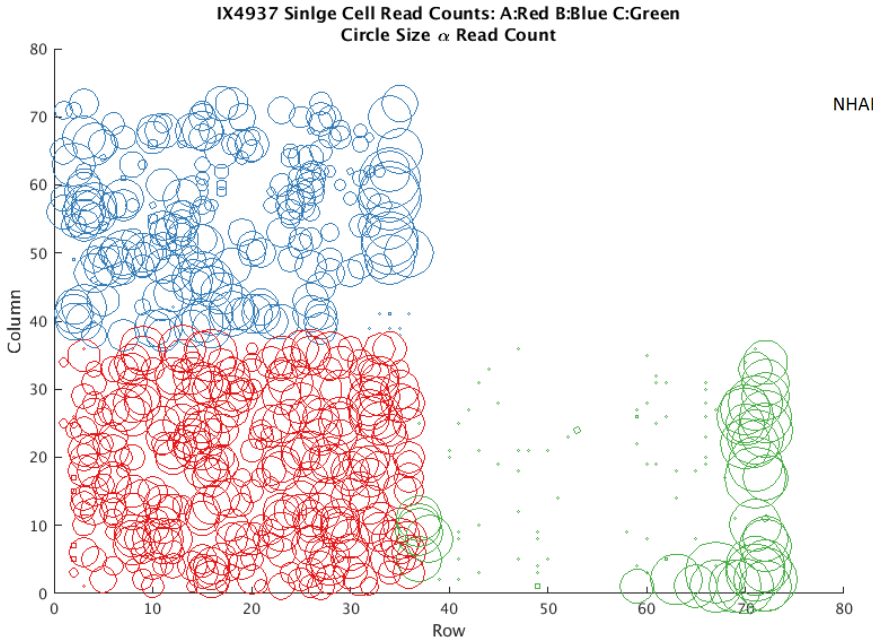
Mate-pairs: large insert, long reads to extend assembly
10X Chromium: Phased genomes, localized assemblies

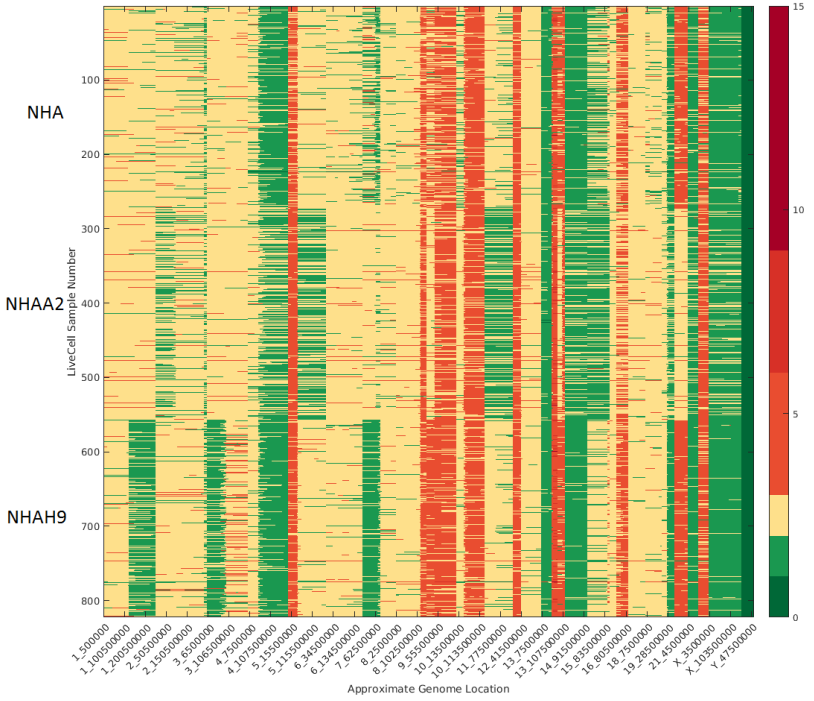
Oxford nanopore: high throughput long reads
Pacbio: consensus long read with lower error rate

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Single-cell sequencing

WGS and RNA sequencing from individual cells allows for single cell resolution of copy number and expression





Cell populations treated under different conditions can be examined separately

Epigenomics

- Post-transcriptional modification cannot be detected through genome and transcritome sequencing
- Efforts such as the International Human Epigenome Consortium have provided comprehensive datasets for comparison and interpretation of epigenomic data
- ChIPseq and bisulphite sequencing (array, whole genome or capture) are used to study histone modification and DNA methylation
- Examples of analysis: Identify genes and pathways that are epigenetically modified, correlated ChIP data with expression and mutational data, cluster samples by DNA methylation profile

Immunogenomics

TCR/BCR sequencing

HLA typing

Analysis from WGS and WTS sequencing:

T and B cell repertoire

HLA typing

Cell type abundance

Neoantigen prediction

How much disk do I need

Data*	Typical file size
30X genome	50G
Full-depth transcriptome	15G
miRNA	500M
1 lane Hiseq 2500	50G
1 lane Hiseq X	65G
Variant files	10-100M

*human data, bam/fastq.gz/raw assembly data are similar in size

Questions

About this talk: yma@bcgsc.ca

About sequencing and bioinformatics at the GSC: dmiller@bcgsc.ca

BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

Acknowledgements

Images and results
Richard Corbett
Reanne Bowlby
Denise Brooks
Emilia Lim
Caralyn Reisle
Eric Chuah
Steve Bilobram

Westgrid
Jana Makar

GSC
Marco Marra, Director
Steven Jones, Director

GSC Group leaders in the room
Andy Mungall, Biospecimen and Library Cores
Richard Moore, Sequencing Core
Robin Coope, Engineering
Diane Miller, Projects

My awesome bioinformatics team
Richard, Nina, Eric, **Kane**, Gina, Irene, Dorothy, Correy, **Dean**, Athena, Dan, Young, Brenna, Laura, **Tina**, Karen, Marc, Tori, Yisu, Patrick, Mya, Darryl, Nav, James, Brandon, **Karen, Morgan**, An, Diana, Johnson, Denise, Shirin, Marcus, **Amir,** Cara, Dustin, Caleb, Daniel, **Simon**, Steve, Reanne, Wei, Sara, Stuart

Lab, Systems, Quality systems, Projects, Admin and all other teams