Towards improved methods for identifying relevant mutations in cancer genomics data

Ryan Morin
Assistant Professor, Department of Molecular Biology and Biochemistry
Simon Fraser University
Scientist, Genome Sciences Centre, BCCA
Faculty, CIHR Bioinformatics Training Program
Outline

• Introduction to cancer genome analysis
  • Limitations and open problems in cancer genome analysis
  • Overview of methods for identifying important genes from mutation data
  • Preliminary results using a new variant calling approach
  • Comparison of performance of tools for identifying cancer genes
  • Lessons learned and future directions
• Other related projects in the Morin lab
Cancer Genome Sequencing*

**Biopsy, Matched Tissue Collection and Pathology**

**Integration and Interpretation**

**DNA/RNA extraction and Sample Prep**

**Sequencing**

Integration and Interpretation

**RB1**

cell proliferation

tumor

Conclusion: A mutation in RB1 caused uncontrolled cell proliferation resulting in the cancerous growth

**Analysis and Genomic Characterization**

**Alignment to Reference Genome**

**Treatment Plan**

**Rx**

Cancer Genome Sequencing*

*Complements of Wikipedia*
Cancer Genomics Consortia

- Hundreds of tumours from common cancer types being studied using NGS

- Analysis can take many directions but consistent overarching goals

  - Changes to gene structure or dosage
    - mutation, expression
  
- Genes recurrently altered in some way to provide the tumour with a fitness advantage

  - Driver genes

- Pathways suppressed or up-regulated

- New/personalized therapeutic options

Human Mol. Gen., Oct 2010
Open Problems

Detection

• Accurate (and sensitive) somatic mutation detection
  • Single Nucleotide Variants (SNVs)
  • Insertion/deletions particularly troublesome
• Copy number variant (CNV) detection
• Structural variation (SV) detection

Interpretation

• Determining significant/relevant targets of mutation
  • Large deletions/amplified regions
  • Many genes altered by point mutations
• Separating passenger mutations from drivers
A Plethora of Options

- Multiple competing tools for sequence alignment/mapping
- Dozens of options for SNV ‘callers’
  - Most expect balance between alleles
- HMM and CBS-based CNV callers
  - Many based on methods for SNP array analysis
  - Often not cancer-specific
- Structural variation (SV) detection
  - Alignment-based: leverage discordant pairs, soft-clipped reads
  - De novo assembly-based approaches

Plenty of room for improvement!
NHL as an example

- Non Hodgkin lymphomas (NHLs) are solid tumour-forming cancers that derive from cells of the immune system
- Collectively among the 5 most common cancer types in adults
- Diffuse large B-cell lymphoma is a common aggressive type of NHL
- 40 DLBCL tumours sequenced locally (GSC) to identify:
  - structural rearrangements
  - CNVs
  - Somatic point mutations*
Confounding Technical Factors

• Data are inherently noisy
• Alignment and variant calling methods are imperfect
• Misalignment can yield artefacts
  • Some are systematic
  • Many appear random
• Many can be recognized by eye
  • Subjective, not scalable
Confounding Biological Factors

• Tumours often contain a substantial portion of normal cells

• Somatic mutations are typically restricted to the the tumour sub-population (not the stroma)

• Not all tumour cells are identical
  - Ongoing mutation and selection can result in multiple related sub-clones

• Both factors can significantly limit our ability to detect somatic mutations

• Driver mutations can arise late during tumour evolution!

Morin et al, 2013; Blood
From Mutation to Biology

- Goal: identify similarities among tumours from many individuals
- Many tumours have high mutation rate/ load
- Larger genes naturally have more mutations due to random acquisition
- Some genes/regions/sequence contexts can be more “mutable” than others
- True signal (driver mutations) are a minority relative to irrelevant mutations and artefacts
- Recurrence can be at the pathway rather than amino acid or gene level
Patterns of mutation recurrence in cancer genes have been shaped by strong selective pressures.

Tumour-specific genome-wide patterns result from distinct mutational processes “mutation spectrum”

Alexandrov et al, 2013; Nature
Hot Spots = Low-hanging Fruit

- Mutation hot spots readily identified -> Potential oncogenes?
- Many genes with mutation hot spots in lymphoma
- Most genes have more diffuse mutation patterns
Patterns of Positive Selection

- Tumours evolve in a Darwinian process
- Cells with mutations that provide a survival advantage are selected for —> Drivers
- Mutations detrimental to cell survival are selected against (cells die/outcompeted)
- Neutral mutations are carried forward during selection —> Passengers
- Genes enriched for drivers should have a higher ratio of non-silent:silent mutations than expected by chance

Ka/Ks < 1  →  Purifying selection
Ka/Ks > 1  →  Positive selection
Strategies for Identifying Driver Genes

- Individual methods leverage different aspects of the data, different types of patterns
- Many aim to determine which genes have more mutations than expected by chance
  - Non-silent vs silent
  - Enrichment for high “functional impact”
  - Significant clustering of mutations in restricted regions of the protein
- Mutations at known regulatory/functional (e.g. phosphorylation) sites may be missed by global selective pressure approaches

Tamborero et al, 2013. Scientific Reports.
Evolutionary Conservation Implies Functional Constraint

Mutations at conserved sites expected to have more potent functional impact
Mutation Load Variable Across 40 Lymphoma Tumours

- Total number of somatic mutations extremely variable
  - Averages ~100 non-silent SNVs
- Single case with no somatic protein-altering changes
  - No drivers?
  - Low tumour content?
- Can increased sensitivity enable detection of somatic variants in such cases?
  - Important for clinical sequencing
Genes with Evidence for Selective Pressure in Lymphoma

- Thousands of somatic single nucleotide variants (SNVs) detected in 40 DLBCL cases
- 74 genes with significant deviation of expected synonymous/non-synonymous mutation ratio
  - Many genes not previously known to be involved in lymphoma
  - Not every genome had a mutation in one of these genes

Morin et al, 2013; Blood
Where might we do better?

- Alignment
  - Improving accuracy of read placement?
- Variant detection
  - Increase sensitivity and specificity
- Identification of functionally relevant genes and pathways
  - Is there an ultimate method to accomplish this?
Improving Accuracy in Detecting Somatic SNVs

• MutationSeq - Machine learning (random forest)-based method for identifying somatic SNVs

• Trained on experimentally-validated somatic SNVs from large number of tumour exomes

• Leverages high-dimensional feature set to reduce false positive rate
  • Recognizes features commonly associated with false positive variant calls

• Tested on 40 DLBCL tumour genomes

Ding et al, 2011; Bioinformatics
MuSeq Increases Specificity

- Repetitive regions of the genome invariably result in enriched false positives in certain genes
  - Combination of misalignment and low-complexity sequence
  - Improved specificity of MutationSeq over standard variant calling approach
  - Known artefacts are reduced but not completely removed
Increased Mutation Yield

- Genomes originally analyzed using early version of MuSeq
- MuSeq3 increased number of somatic SNVs detected in almost all cases
  - Some sites visually convincing but need experimental confirmation
  - Include known polymorphisms (likely under-sampled SNPs)
- Substantially improved yield in single problematic low-tumour content case
Improved Sensitivity in Suspected Cancer Genes

RNA-seq

Genome

Somatic variant
Resolution of Sub-clonal Variants?
Leveraging Additional DLBCLs

- Three groups have performed large-scale sequencing in diffuse large B-cell lymphoma
  - 40 cases (Vancouver cohort - Morin et al, 2013)
  - 49 cases (Boston cohort - Lohr et al, 2011)
  - 36 cases (North Carolina cohort - Zhang et al, 2013)
- Pool data and perform meta-analysis using MuSeq3
- Analyze for significant genes using a variety of strategies
- Did any of these three studies miss any important genes?
Have we just hit the tip of the Iceberg?
Or does this result from differing criteria for significance?

*Morin et al., 2011; Nature. Actually only reported 26 significant genes*
Overview of SNV calls

• 121 cases suitable for paired analysis
• 23,526 Missense variants
• 12,950 Silent variants
• 1300 Nonsense variants
• 638 splice site mutations
• 36 stop-lost variants
Significance Methods Tested

- MuSiC - Significantly mutated genes (correct for gene length, sequence coverage)
- MutSigCV - Significantly mutated genes and correcting for correlates suspected to impact local mutation rate
- OncodriveFM - Enrichment of mutations at conserved sites
- OncodriveClust - Clustering of mutations within the linear protein sequence
- ActiveDriver - Enrichment of mutations at phosphorylation sites
MuSiC: Too Inclusive?

• 292 significantly mutated genes (FDR = 0.05)
  • Includes 27 of 74 significant genes identified in smaller (local) cohort
• Many genes of unknown relevance/suspected irrelevance
• Also many genes in pathways known to be important in lymphoma and other cancers
  • lymphocyte differentiation
  • Antigen presentation (MHC)
  • apoptosis
• Long tail of infrequently-mutated driver genes?
MutSigCV: Too Conservative?

- Treats average gene expression level and replication “timing” of gene as covariates.
- Few significant genes detected (FDR = 0.2).
  - Fewer than reported in original (Lohr) study.
- Many well known genes were below threshold.
- P-value over-correction due to many more tests?
- Is inclusion of sub-clonal mutations (and more artefacts) reducing statistical power?
Genes with Mutation Clusters

- OncodriveClust attempts to identify significant clusters of mutations within proteins
- More sensitive method for detecting rarely occurring activating mutations?
- Some known lymphoma-related hot spots identified: EZH2, CD79B, STAT6
- Many additional mutation clusters identified appear to be recurrent artefacts
  - Example: PTPN11 recurrent nonsense mutation
  - Attributable to read mapping error/ambiguity
Significantly Mutated Phosphorylation Sites

- ActiveDriver identifies SNVs that modify any of 73,872 experimentally determined phosphorylation sites

- 100 genes with 1+ mutated phosphorylation site (pSNV)
  - Either Y/S/T residue or adjacent recognition motif altered

- Small overlap with MuSiC results

- Include known oncogenes and tumour suppressors acting in lymphoma
  - TP53, ETV1, BCL2, CD79B, CARD11

- Three mutations predicted to affect phosphorylation site in BRAF oncogene

- Recurrent mutations in PIK3R1 (D560G), previously reported to be activating in glioblastoma
Recurrent Artefacts are Pervasive and Problematic

- Many clustered mutations and phosphorylation site mutations (pSNVs) appear dubious on visual inspection
- Often attributable to misalignment
  - Proximal insertion/deletion variants
  - Gene families, un-annotated pseudogenes, low-complexity regions
- Misalignment results in consistent variants in multiple samples
  - Not readily recognizable as problematic by variant callers
  - Implicitly using recurrence in un-related samples?
- Improve the read alignment process?
- Selecting a set for experimental confirmation and re-training
Lessons Learned

• Many approaches for identifying significant mutation patterns exist

• There is no “one” tool that can identify all relevant genes

• Complementary methods leverage different aspects of the data

• Application to large NHL meta-analysis identified many (or few) significant genes, depending on method

• Seemingly poor consensus between approaches

• New genes of interest have emerged that were not identified in individual DLBCL sequencing studies

• Recurrence-based approaches impacted by artefacts
Can we Improve Mutation Cluster Detection?

- Clustering based on position in amino acid sequence is limited
- Amino acids distal in two dimensions can be proximal in three-dimensional space
  - Enzyme active site
  - Protein-protein interaction interfaces
- Three dimensional clustering of mutations would not necessarily be detected as hot spots
- Can 3D clustering of mutations be automatically detected?

Lymphoma mutations in GNAI2 mapped to GNAI3 structure
SpacePAC: 3D mutation clustering

- Attempt to overlay spheres on 3D protein structure to encapsulate mutations
- Identifies optimal set of spheres (minimal size/#) to capture the maximum number of mutations
- Computes P value for mutations randomly clustering in set of spheres
- Only applicable for proteins with solved structure
  - Or homology modelling?
  - Computationally demanding
  - Not yet applied to lymphoma meta-analysis

How can this help cancer treatment in the short term?

- Patients receiving standard therapies continue to succumb to cancers
- Real-time studies are testing utility of clinical genome sequencing to guide treatment
- How can we leverage prior knowledge to maximize utility of clinical sequencing?
  - Improve mutation detection?
  - Identify clinically-relevant mutations?

*Exact sensitivity of genome sequencing not actually known
**Determining functional variants when n=1 is an open problem
Digging Deeper into Tumour Genomes

- Goal: maximize recall of driver mutations from each tumour
- Focus in-depth analysis on restricted set of genes with potential significance
- Use superior alignment methods to more accurately place reads
  - Remove recurrent artefacts
  - Improve detection accuracy for indels
  - Re-call mutations with more lenient thresholds
- Plan to test method on related cancer types
- May facilitate detection of more sub-clonal variants or “actionable” mutations in clinical setting

Diagram:
- Tumour
  - Somatic mutation detection*
    - Functional annotation**
      - Significance/recurrence testing
    - Reduced stringency variant calling
      - Driver mutations/genes
        - Targeted re-alignment
          - n=100+
          - Insight into disease biology
Liquid Biopsies to Monitor Tumour Evolution

- Solid tumours (including lymphomas) release genomic DNA into circulation as circulating tumour DNA (ctDNA)

- Tumour-derived cell-free DNA can be differentiated from normal DNA by presence of:
  - CNVs, SNVs, Structural rearrangements

- Mutations can be detected in plasma using sensitive personalized assays
  - New non-invasive tool for monitoring disease course?
  - Can monitor/compare levels of individual mutations
  - Potential for detecting mutations *en masse* using global sequencing (e.g. exome)
  - Limited to cases with sufficiently high ctDNA levels
Towards Plug and Play Cancer Genome Analysis

- Galaxy is an open source, web-based platform for data intensive biomedical research
  - Facilitates the cloud-based application of command-line bioinformatics software
  - Graphical interface allows individual tools and workflows to be run without knowledge of shell or high-performance computing
Galaxy Workflows for Cancer Genome Analysis

- Implementing open-source locally-developed tools or other utilities useful for cancer genome analysis in Galaxy framework

- Deployed in virtual machine, ultimately portable to Amazon EC2 or other clouds

- Default workflows readily user-modifiable to swap in additional tools

- Allow processing of single samples and leveraging mutational data from related cancers

- Ultimately aim to include:
  - integrative methods for identifying driver mutations
  - Tools for visualization
  - Report of known/actionable somatic mutations

MutationSeq
galaxy tool

Somatic SNV workflow
Canine Lymphoma as a Model for Human Cancer

- Cancer is the most common cause of premature death in dogs
- NHLs represent 20-25% of all canine cancer cases
  - Similar histology to human DLBCL
- Rate of cancer in dogs is much higher than in humans
- Treated with similar therapies
- Are the genes involved in human NHL also involved in the canine form of this cancer?
  - Comparative cancer genomics
- Potential opportunity for testing new prognostic markers and therapies?
Acknowledgements

Morin Lab
- Dan Fornika
- Bruno Grande
- Divesh Mestry
- Kevin Bushell
- Stephen Yu

Marra Lab
- Marco Marra
- Rodrigo Goya
- Diane Trinh

Collaborators
- Sohrab Shah & Shah lab
- Paul Boutros
- Randy Gascoyne
- Joe Connors
- Nathalie Johnson

VanBUG organizers!