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

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Detecting structural variants

February 14th 2018

Karen Mungall, Bioinformatics Coordinator - Analysis Groups
Canada's Michael Smith Genome Sciences Centre
BC Cancer Agency

Talk overview

- Part 1
  - What is a gene fusion
  - How do they arise
  - Why are they important
- Part 2
  - Considerations for tool selection
  - What to do with the data
- Part3
  - Comprehensive SV detection

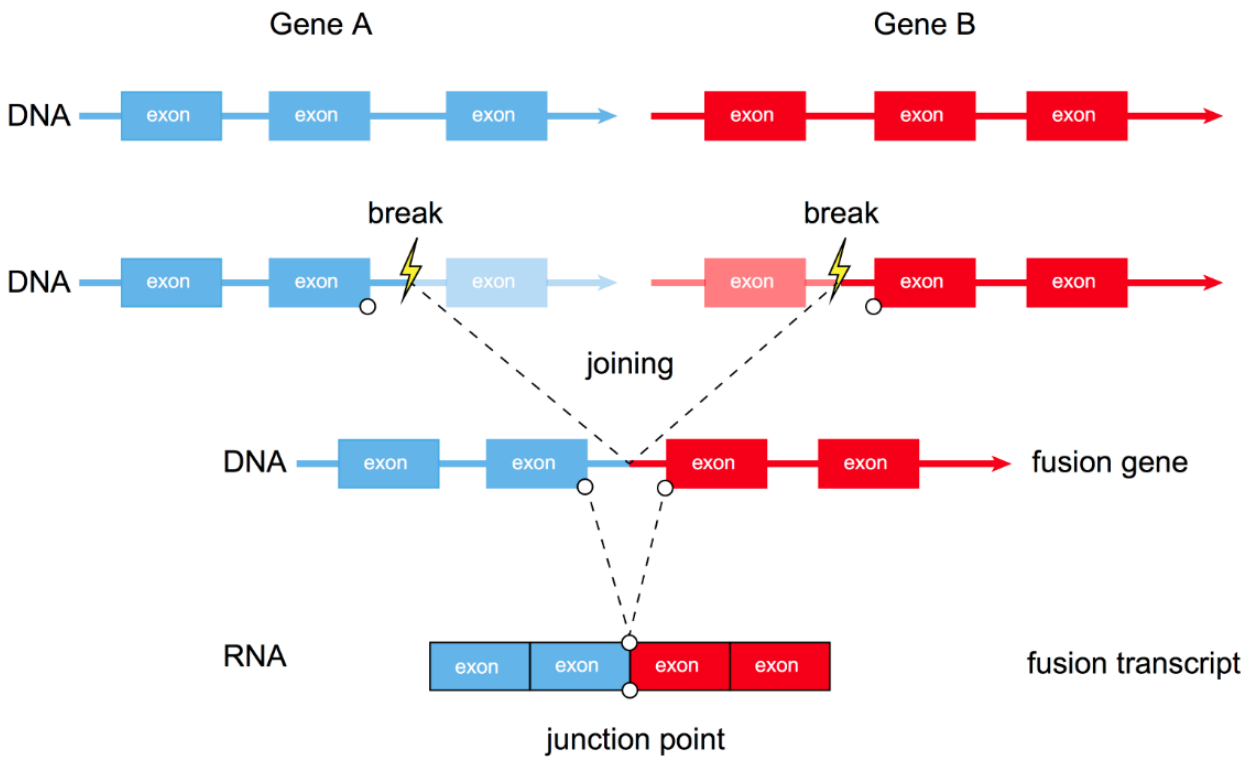
PART 1

PART 1

Gene fusions - What, how, why?

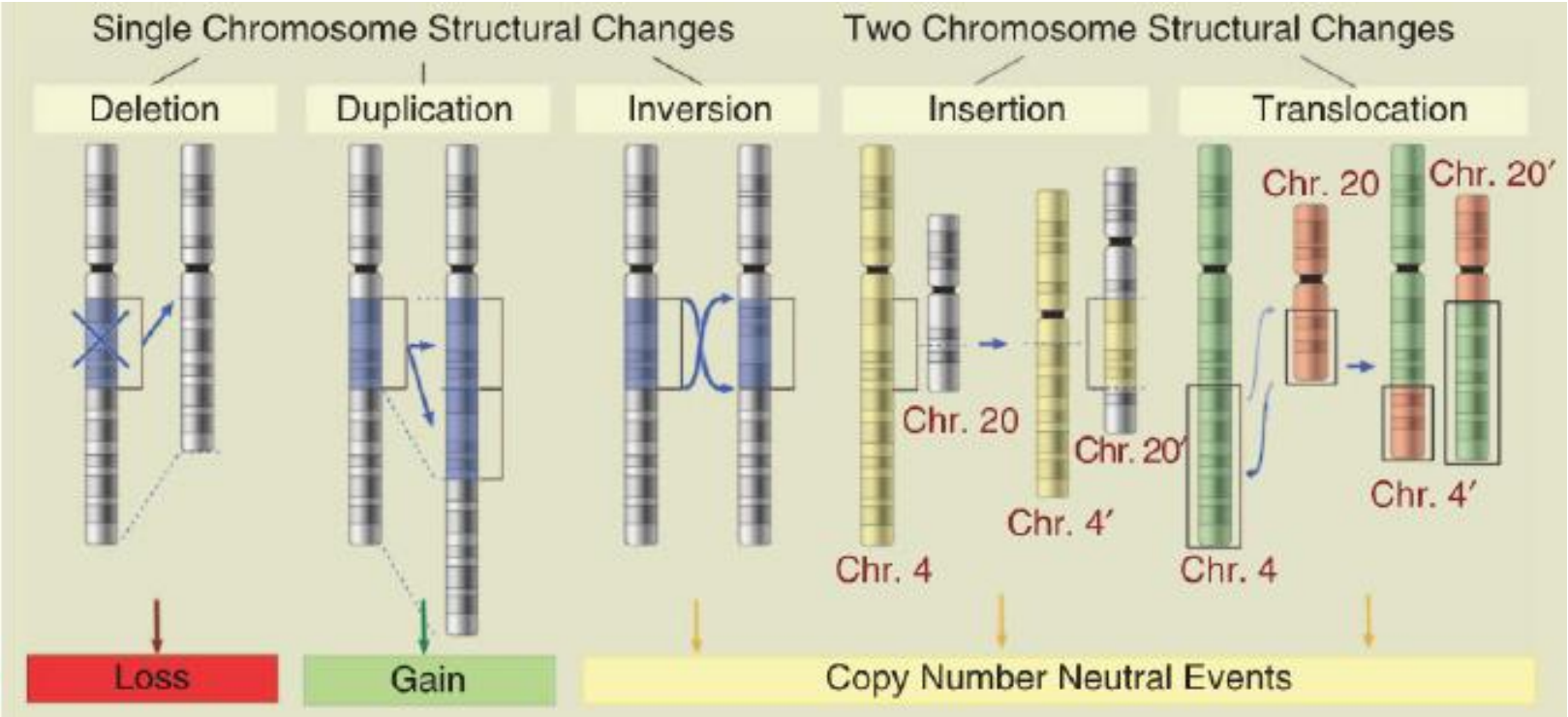
What is a gene fusion?

- When two separate genes come together to form a new chimeric gene. The resulting protein product may lead to abnormal expression levels and function and may in turn cause the abnormal proliferation of cells and cancer development



http://www.tumorfusions.org/

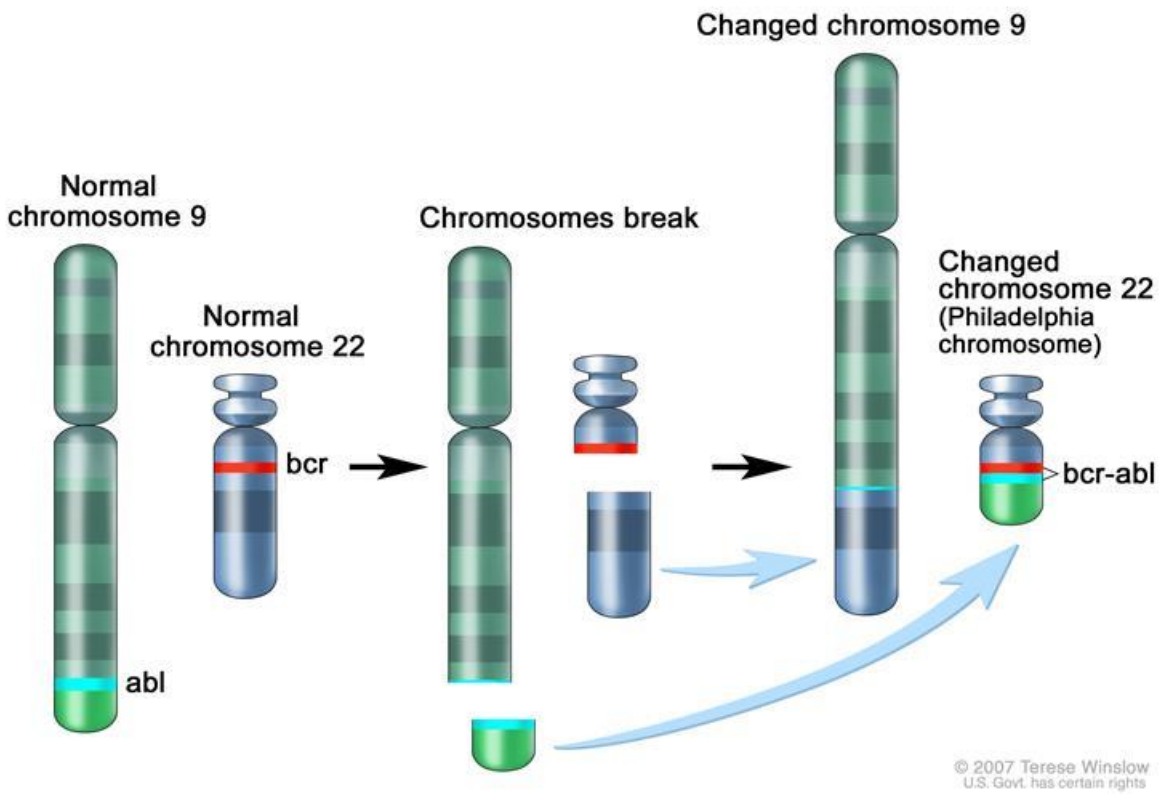
How does a gene fusion arise?



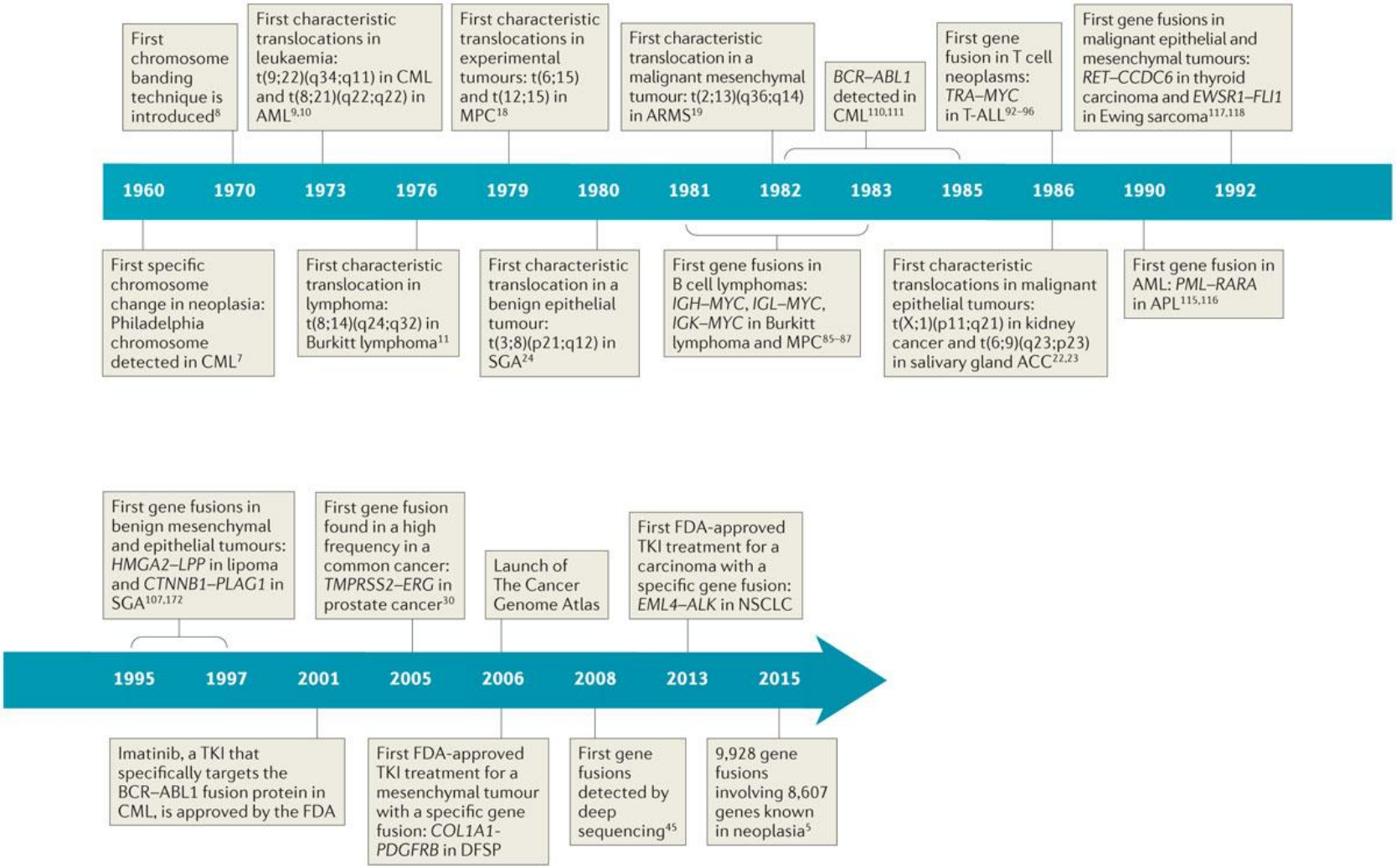
Recent Advances in Autism Spectrum disorder, Edition: 1, Chapter: 16, Publisher: InTech, Editors: Prof. Michael Fitzgerald, pp.341-358

The first gene fusion

The first fusion gene identified, is known as the Philadelphia chromosome. It arises from a translocation event involving the 5'part of the BCR gene on chr22 fusing to the 3'part of the ABL1 gene on chr9. It was first discovered in chronic myelogenous leukemia (CML). *BCR-ABL1* has been found to occur in more than 95% of CML patients and to exert its oncogenic phenotype by encoding a constitutively active ABL1 kinase

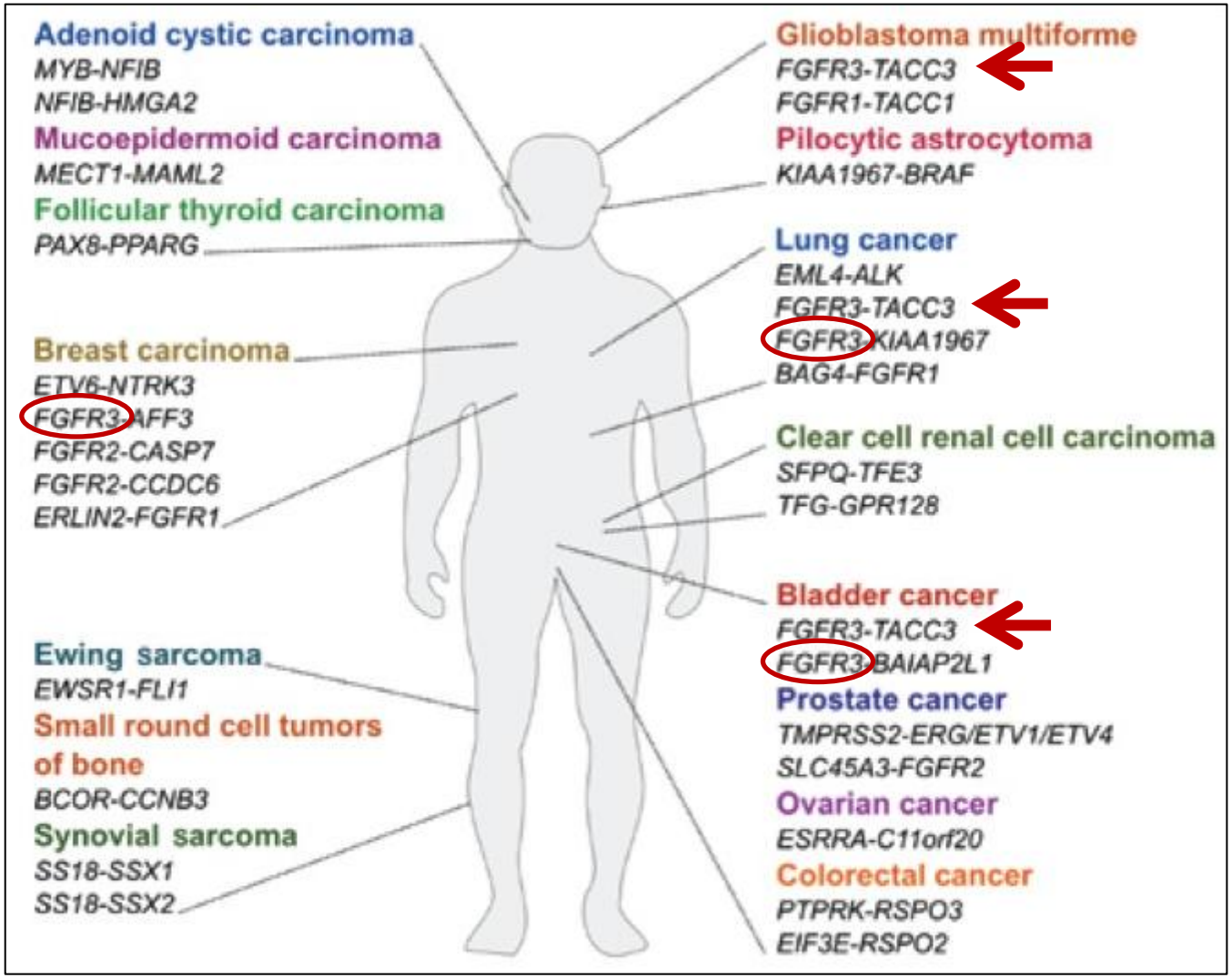


Gene fusion time line



Nature Reviews Cancer 15, 371–381 (2015**) The emerging complexity of gene fusions in cancer**

Recurrency in gene fusions

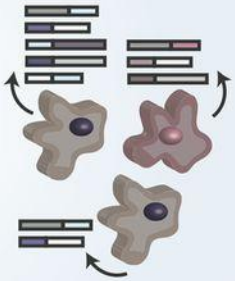


Chin J Cancer. 2013 Nov; 32(11): 594–603.**Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment**

Trends in fusion functionality

A Gene fusion landscapes are diverse

The diversity, abundance, and connection to etiology of gene fusions varies across both cancers and individuals

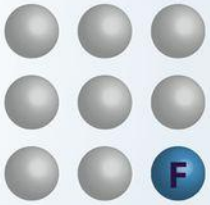


B Gene fusion networks elucidate fusion pairings

Network studies show that most fusion genes fuse with very few partners, and that different cancer types have signature fusion networks

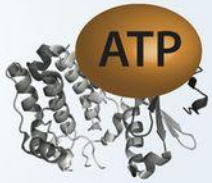
C The frequency of fusions in cancers varies considerably

Fusions tend to be rare, but can be predominant, and anti-correlate with other somatic mutations



D Fusion genes tend to have specific functions

Molecular functions relating to kinase or DNA-binding activity are enriched in genes forming fusions



Nucleic Acids Res. 2016 Jun 2; 44(10): 4487–4503. **Discovering and understanding oncogenic gene fusions through data intensive computational approaches**

Structural features of fusion proteins

A Breakpoint locations tend to preserve protein function

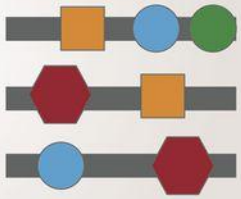
Breakpoints tend to occur in disordered regions and maintain reading frames and protein globularity

B Fusion proteins are relatively depleted in domains

Proteins which form fusions have fewer domains than other proteins, but fusion transcripts encode more domains than expected by chance.

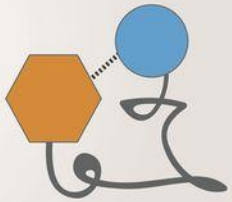
C Fusion proteins contain specific domain architectures

Domain recombinations in fusion proteins are non-random and sometimes novel



D Disorder may contribute to fusion protein functionality

The increased disorder in fusion proteins could promote the viable joining of different domains and offer flexibility for internal interactions



Nucleic Acids Res. 2016 Jun 2; 44(10): 4487–4503. **Discovering and understanding oncogenic gene fusions through data intensive computational approaches**

Prognostic significance

- **TMPRSS2-ERG Fusion Gene in Prostate Cancer**

High expression of TMPRSS2-ERG gene fusion together with prostate-specific antigen levels are indicators for likelihood of recurrence and shortened time to recurrence

Prognostic Significance of *TMPRSS2-ERG* Fusion Gene in Prostate Cancer

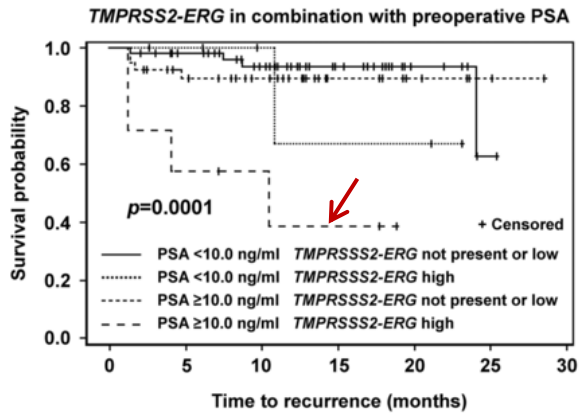


Figure 2.

Relation of TMPRSS2–ERG fusion transcript expression in combination with preoperative serum PSA level to time to recurrence (Kaplan–Meier curves). A combination of high PSA level and high TMPRSS2–ERG expression was associated with the shortest time to recurrence.

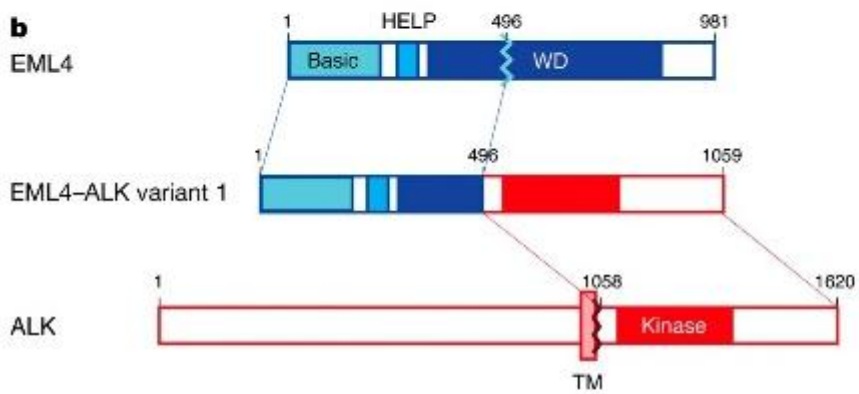
Anticancer Res. 2016 Sep;36(9):4787-93. **Prognostic Significance of TMPRSS2-ERG Fusion Gene in Prostate Cancer.**

Diagnostic significance

- Confirmation of diagnosis
  - BCR-ABL1 in CML patients hall mark fusion seen in ~95% patients
- Specific subgroup:
  - EML4-ALK fusion is seen in around 5% of NSCLC patients

Therapeutic significance

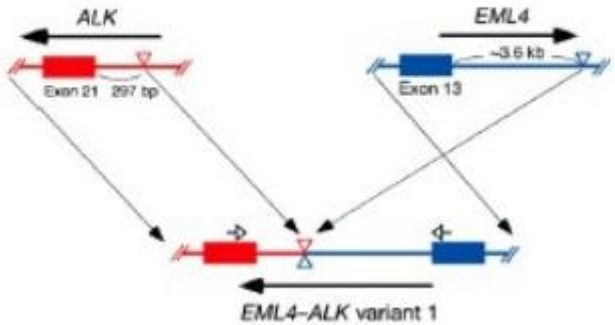
EML4-ALK



Fusion of the N-terminal EML-4
(the basic region, the HELP domain and part of the WD repeat region)

to the

Intracellular region of ALK
(the tyrosine kinase domain)



Both EML4 and ALK genes map to short arm of chromosome 2p, with opposite orientations

Soda M; Nature; 2007

Therapeutic

- Patients with *ALK* rearrangements **do not** benefit from EGFR-specific TKI therapy but may be considered for therapy targeting the constitutively activated receptor tyrosine kinase that results from *EML4-ALK* and other *ALK* fusions. Crizotinib is the first FDA-approved ALK TKI. It is indicated for treatment of locally advanced or metastatic NSCLC in patients whose tumors are positive for ALK as determined using an FDA-approved test.

- Additionally, *EGFR*, *KRAS*, and *ALK* mutations are almost always mutually exclusive (ie, mutations of only 1 of the 3 genes occur within any individual tumor).

- Methods for detecting the *ALK* rearrangements include FISH, PCR, and immunohistochemical (IHC) staining. ALK tests are often run in conjunction with tests for EGFR and KRAS mutations

- Outcome: Sensitive to ALK inhibitors eg Crizotinib Resistant to EGFR Tyrosine Kinase Inhibitors

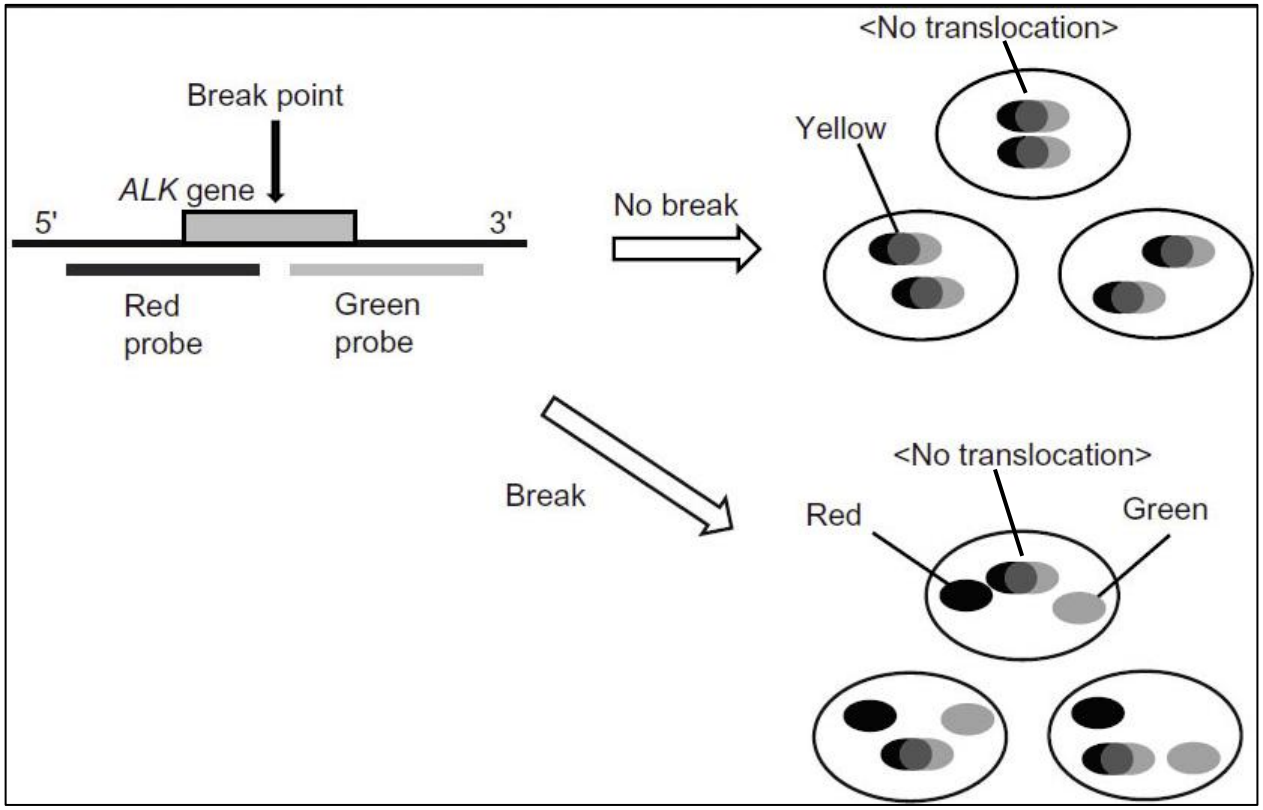
PART 2

PART 2

Tool selection and what to do with results

How to detect fusions

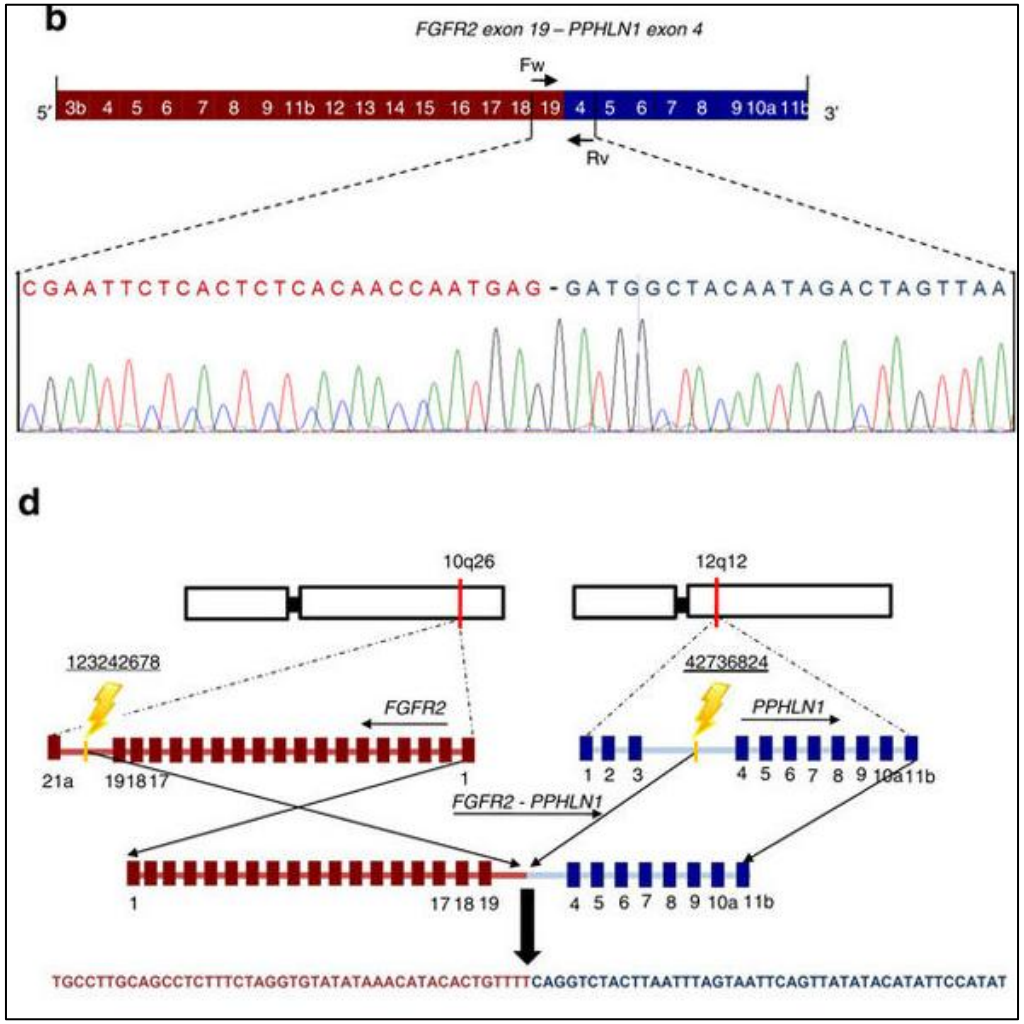
Fluorescent in situ Hybridization (FISH)



Onco Targets Ther. 2014 Mar 5;7:375-85 **Development of anaplastic lymphoma kinase (ALK) inhibitors and molecular diagnosis in ALK rearrangement-positive lung cancer.**

How to detect fusions

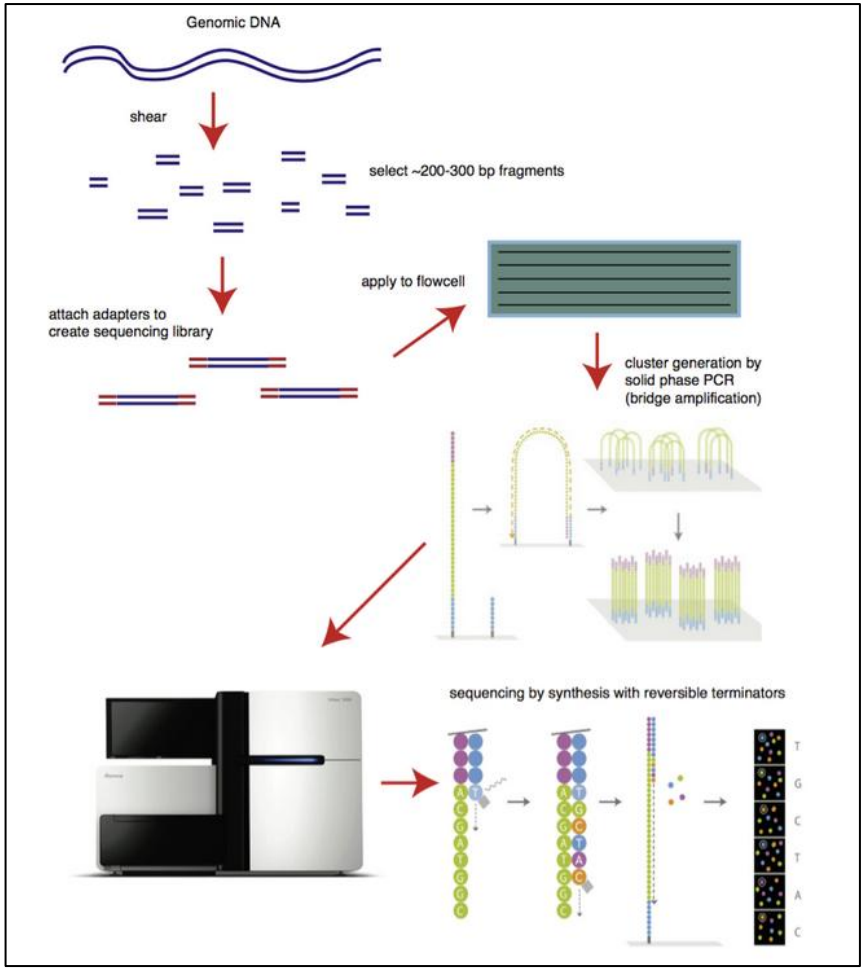
Polymerase Chain Reaction (PCR)



Nat Commun. 2015 Jan 22;6:6087 **Massive parallel sequencing uncovers actionable FGFR2-PPHLN1 fusion and ARAF mutations in intrahepatic cholangiocarcinoma.**

How to detect fusions

Massively parallel sequencing (Illumina sequencing by synthesis)



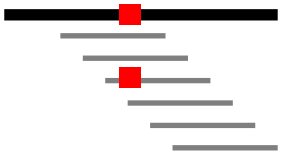
https://bitesizebio.com/

Targeted fusion detection

- Advantages
  - specific and fast
- Disadvantages
  - need to know ahead of time what you want to find

Identify actionable event
(can be SNV, indel, fusion breakpoint, fusion gene partners or exon in gene)

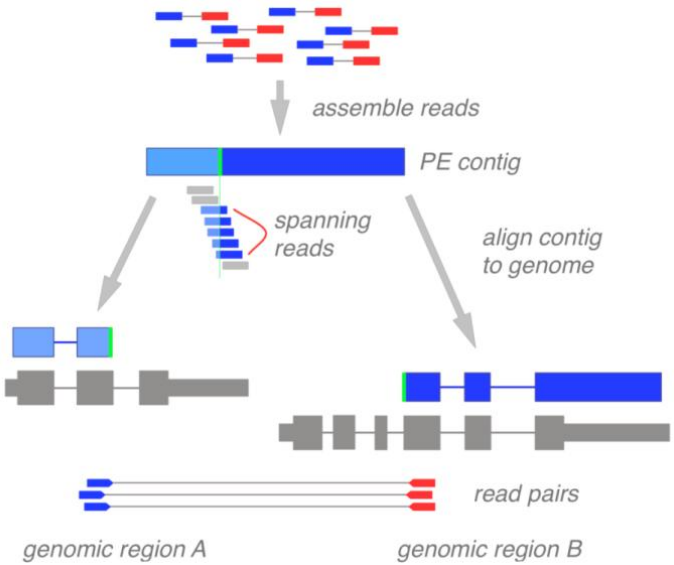
Create probe sequence containing event



Screen probe sequence with target fastq or bam file

De novo fusion detection

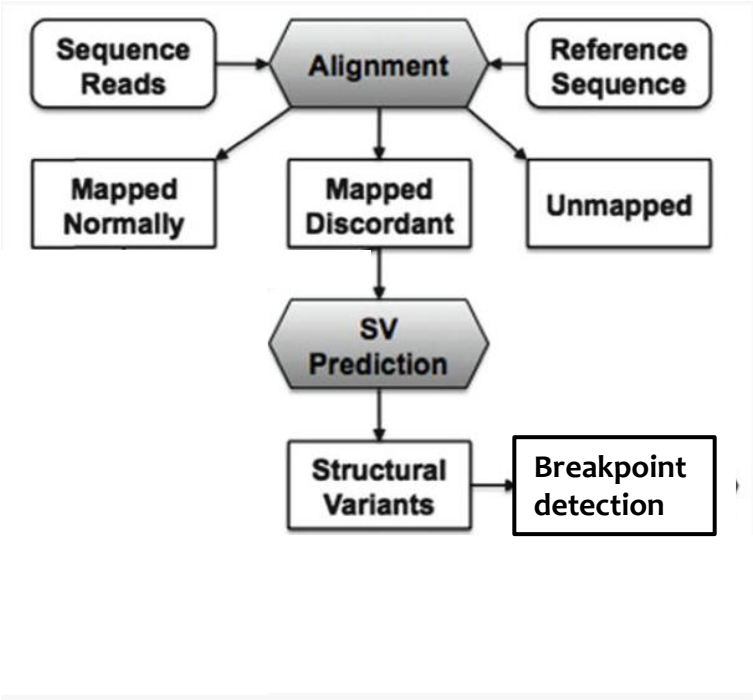
- Assembly based fusion detection
  - Advantages
    - Comprehensive event detection
    - Higher specificity
    - generates contigs for better interrogation of event breakpoint
  - Disadvantages
    - Large resource requirement with multiple steps and slow
    - Lower sensitivity



Nat Methods. 2010 Nov;7(11):909-12. **De novo assembly and analysis of RNA-seq data**

De novo fusion detection

- Alignment based fusion detection
  - Advantages
    - Fast with lower resource requirement
    - Higher sensitivity
  - Disadvantages
    - lower specificity



Adapted from Methods Mol Biol. 2012; 838: 369–384. **Massively Parallel Sequencing Approaches for Characterization of Structural Variation**

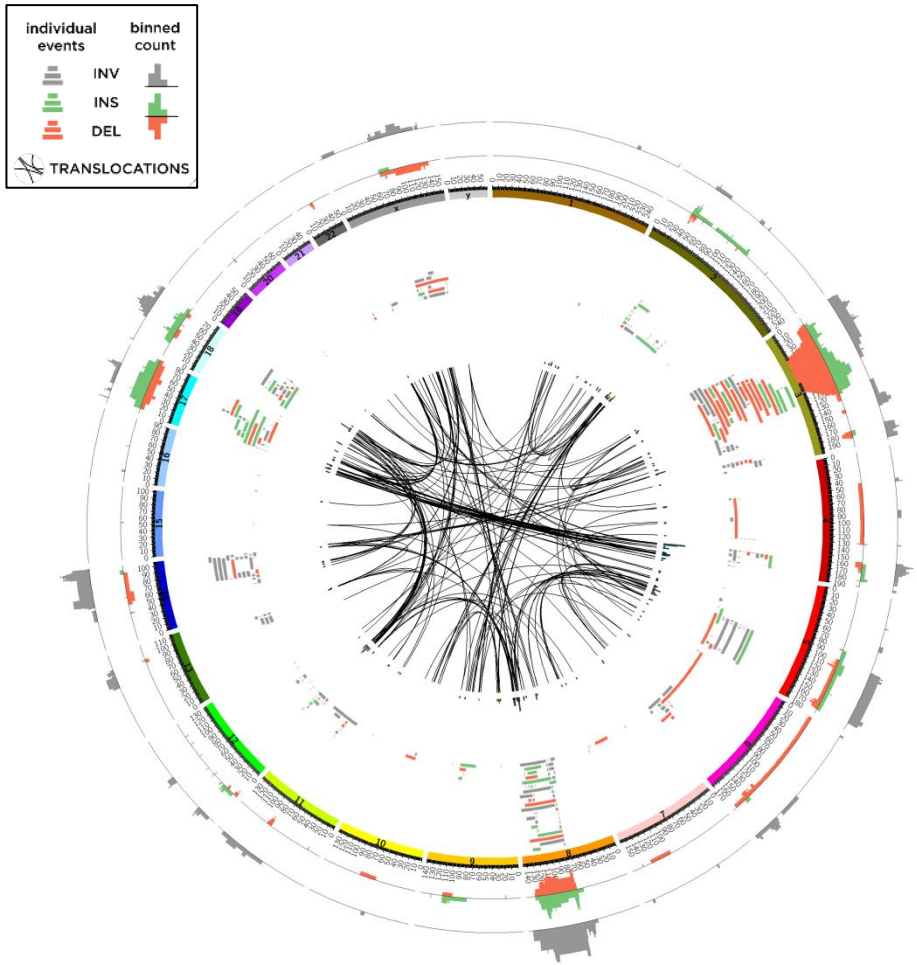
Considerations for tool selection

- Sample source
  - Fresh or FFPE
- Data type
  - DNA or RNA
- Input data
  - Fastq or bam
  - readlength
- Test samples
  - Individual matched samples eg DNA, RNA
  - Multiple individual analysis eg trio

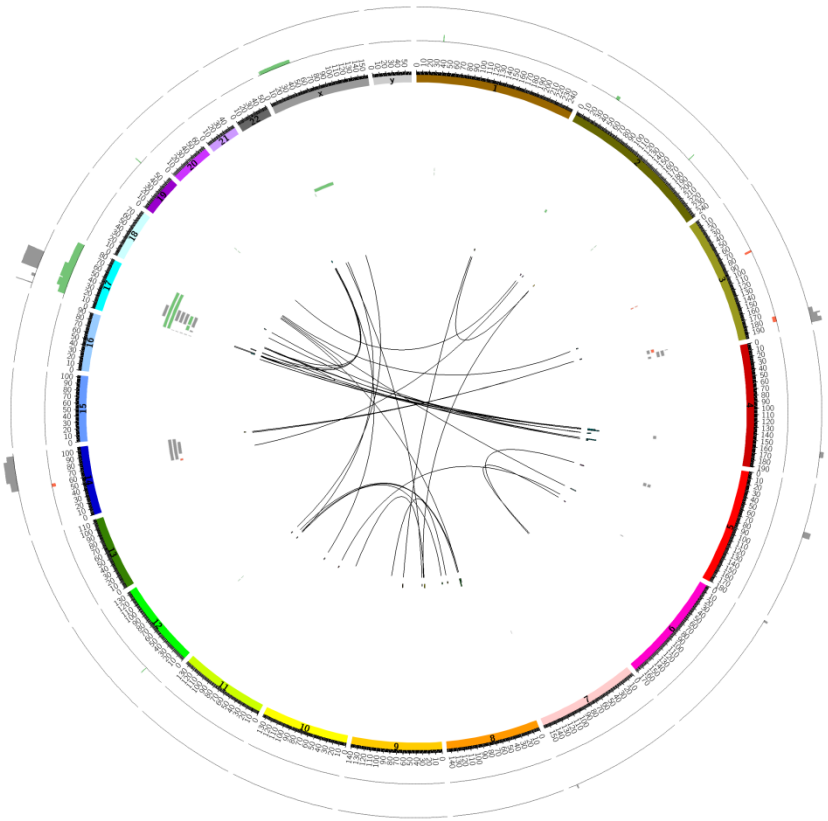
Other tool considerations

- Sensitivity
- Specificity
- Speed
- Resources
- Deterministic

Visualizing data



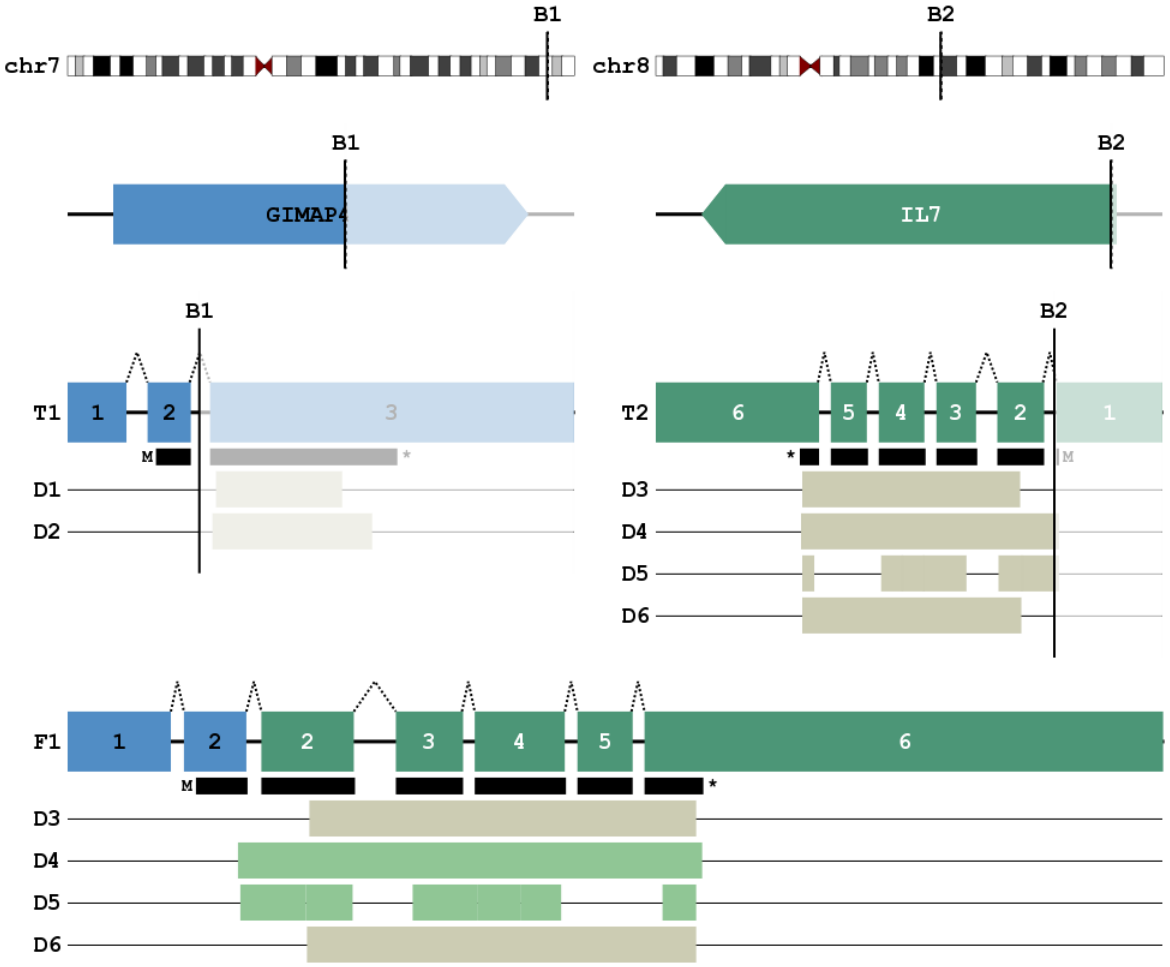
Genome



Transcriptome

Genome Res. 2009. 19:1639-1645 **Circos: An information aesthetic for comparative genomics**

Gene fusion visualization



Bioinformatics – submitted **MAVIS: Merging, Annotation, Validation, and Illustration of Structural variants**

Comparison with tumour data sets

- http://www.tumorfusions.org/
- COSMIC
- Mitleman

Tumor Fusion Gene Data Portal

The Jackson Laboratory

TUMOR FUSION GENE DATA PORTAL

Navigation
Introduction
Summary
By Gene
By Fusion
By Sample
Cancer Type
Normal Tissue
Download
Help
Contact Us

Landscape of cancer-associated fusions using the Pipeline for RNA sequencing Data Analysis

Transcripts fusion as a result of genomic rearrangement is an important class of somatic alteration, as a cancer initiating event and as a molecular therapeutic target for specific tumors. Our Pipeline for RNA sequencing Data Analysis (PRADA) enables us to detect fusion transcripts with high confidence comprehensively. Based on integrated analysis of paired-end RNA sequencing and DNA copy number data from The Cancer Genome Atlas(TCGA), The Tumor Fusion Gene Data Portal provides a bona-fide fusion list across many tumor types.

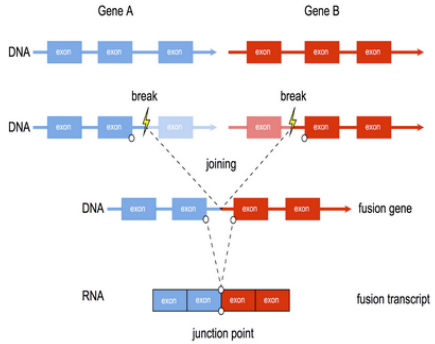


Figure 1. Fusion transcripts. Fusion transcripts are chimeric mRNAs encoded from the joined parts of two genes, and may occur as a result of genomic rearrangements.

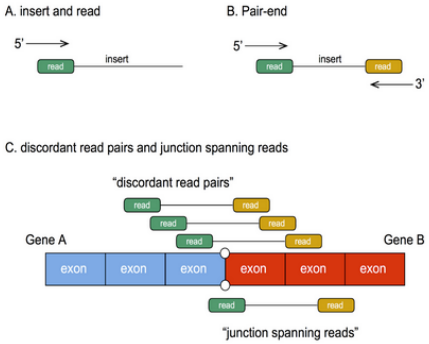


Figure 2. Detection of fusion transcripts. PRADA detects fusion transcripts through identification of discordant read pairs and junction spanning reads.

http://www.tumorfusions.org/

Catalogue Of Somatic Mutations In Cancer

COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects | Data | Tools | News | Help | About | Genome Version | Search COSMIC... SEAR

COSMIC v83, released 07-NOV-17

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

COSMIC
The core of COSMIC, an expert-curated database of somatic mutations

Cell Lines Project
Mutation profiles of over 1,000 cell lines used in cancer research

COSMIC-3D
An interactive view of cancer mutations in the context of 3D structures

Cancer Gene Census
A catalogue of genes with mutations that are causally implicated in cancer

Data curation

- Gene Curation — details of our manual curation process
- Gene Fusion Curation — details of our curation process for gene fusions
- Genome Annotation — information on the annotation of genomes
- Drug Resistance — curation of mutations conferring drug resistance

http://cancer.sanger.ac.uk/cosmic

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer

National Cancer Institute — U.S. National Institutes of Health | www.cancer.gov

CANCER GENOME ANATOMY PROJECT

CGAP How To | Genes | Chromosomes | Tissues | SAGE Genie | RNAi | Pathways | Tools

Chromosomes

Tools
- FISH-mapped BACs
- Genetic and Physical SNP Maps
- Mitelman Searchers
- SNP500Cancer

CGAP Data
- Download

Purchase CGAP Reagents
- CCAP BAC Clones

Related Links
- Atlas of Genetics & Cytogenetics in Oncology & Haematology
- Progenetix
- SKY/M-FISH & CGH Database

Quick Links:
- ICG
- NCI Home
- NCICB Home
- NCBI Home
- OCG

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer

Searching the Database

The information in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer relates chromosomal aberrations to tumor characteristics, based either on individual cases or associations. All the data have been manually culled from the literature by Felix Mitelman, Bertil Johansson, and Fredrik Mertens.
CGAP has developed six web search tools to help you analyze the information within the Mitelman Database:

- The Cases Quick Searcher allows you to query the individual patient cases using the four major fields: aberration, breakpoint, morphology, and topography.
- The Cases Full Searcher permits a more detailed search of the same individual patient cases as above, by including more cytogenetic field choices and adding search fields for patient characteristics and references.
- The Molecular Biology Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to gene rearrangements as a consequence of cytogenetic aberrations.
- The Clinical Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to clinical associations of cytogenetic aberrations and/or gene rearrangements.
- The Recurrent Chromosome Aberrations Searcher provides a way to search for structural and numerical abnormalities that are recurrent, i.e., present in two or more cases with the same morphology and topography.
- The Reference Searcher queries only the references themselves, i.e., the references from the individual cases and the molecular biology and clinical associations.

Database last updated on December 14, 2017
Total number of cases = **68,170**
Total number of gene fusions = **11,124**

Need help! To learn about the Mitelman Database and how to search it, please visit:
- All about the Mitelman Database, which provides background information about the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.
- Mitelman Database Search Help, which contains information on how to use the search tools.
- ISCN Abbreviated Terms and Symbols, which provides a list of terms and symbols used to describe chromosome abnormalities.

Citation of the Database

https://cgap.nci.nih.gov/Chromosomes/Mitelman

PART 3

PART 3

Comprehensive structural variant detection

Comprehensive structural variant detection

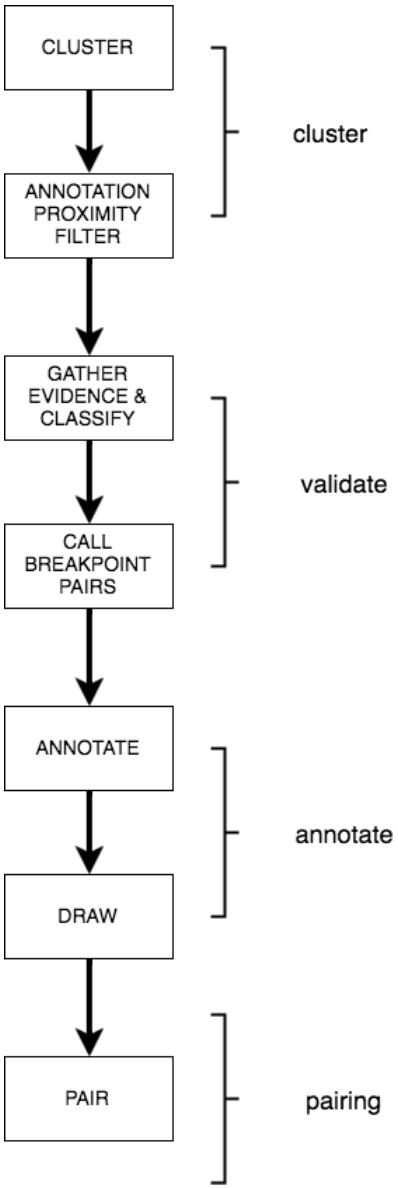
- Multiple tool input
- Clustering of breakpoints
- Consistent breakpoint calling
- Data pairing
- Evidence support
- Standard annotation
- Standard output format

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

MAVIS

Merging, Annotation, Validation, and Illustration of Structural variants

MAVIS process outline



1. Cluster
2. Filter based on proximity to annotations
   - Call a new merged breakpoint pair from the group
3. Gather read evidence
4. Call breakpoint pairs (contig, split read, flanking pairs)
   - contig, split read, flanking pairs
5. Annotate with gene and transcript level information
   - Build Fusion Transcripts for exact calls
6. Draw SVGs for all calls
7. Pair calls between libraries
   - Somatic, Expressed
8. Summary
   - Standard output file with HGVS nomenclature

Merging

- Takes inputs from any SV caller as long as it is put in a common format
- filters based on user-defined masked regions
- Splits calls by type
- Merges based on proximity
  - uses a clique finding algorithm
  - followed by hierarchical clustering for larger clusters than cannot be computed exactly inexpensively

Validation

- Uses bam files to collect support for the input event calls
  - Uses read pair fragment distribution to define intervals of where reads will be collected from
  - collects spanning, split, and half-mapped reads
  - collects flanking and compatible-flanking pairs
  - standardizes cigar/read-alignments to ensure reproducible calls
  - uses a collapsed annotation model to adjust these intervals and calculate read-pair fragment sizes for transcriptomes
- Local assembly
  - Does not attempt to resolve or assemble repeats longer than the read/kmer length
- Calling breakpoint pairs
  - call by contig
  - call by split reads when calling by contig fails
  - call by flanking (or split and flanking) pairs when calling by split reads fails

Annotation

Gene level
Nearby genes
Genes encompassed by the event
Genes at the breakpoints

Fusion Transcript level
Exon/intron of breakpoint
Uses a splicing model to predict if the fusion will be in or out of frame
Predicts domain retention by re-aligning protein domain sequences to the new amino acid sequence of the fusion protein

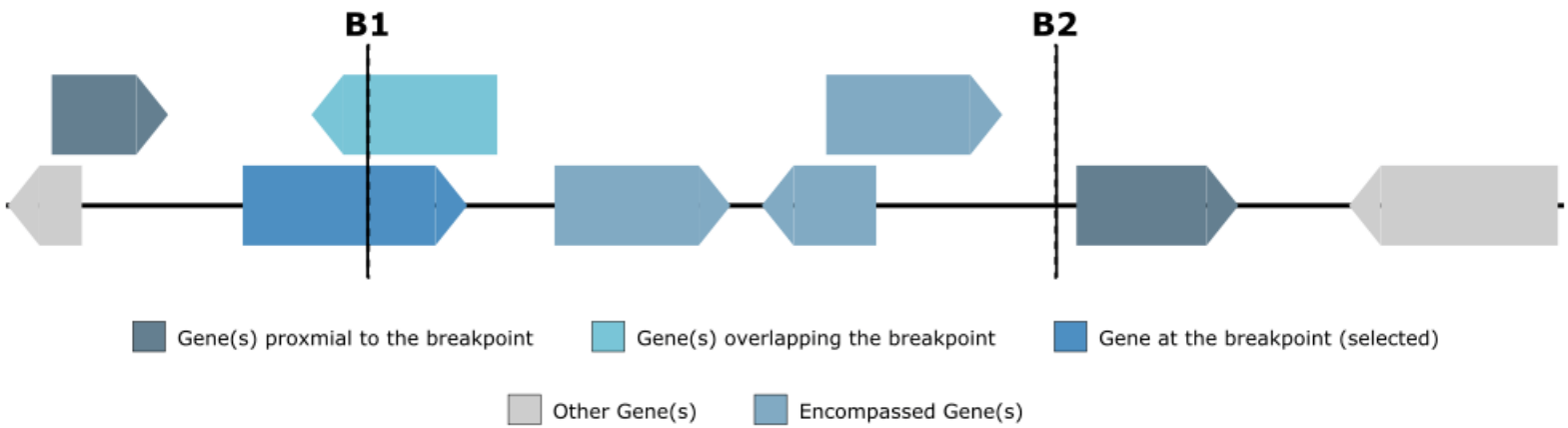


Illustration- gene fusion

For any fusions with breakpoint level resolution a figure and putative splicing products are produced

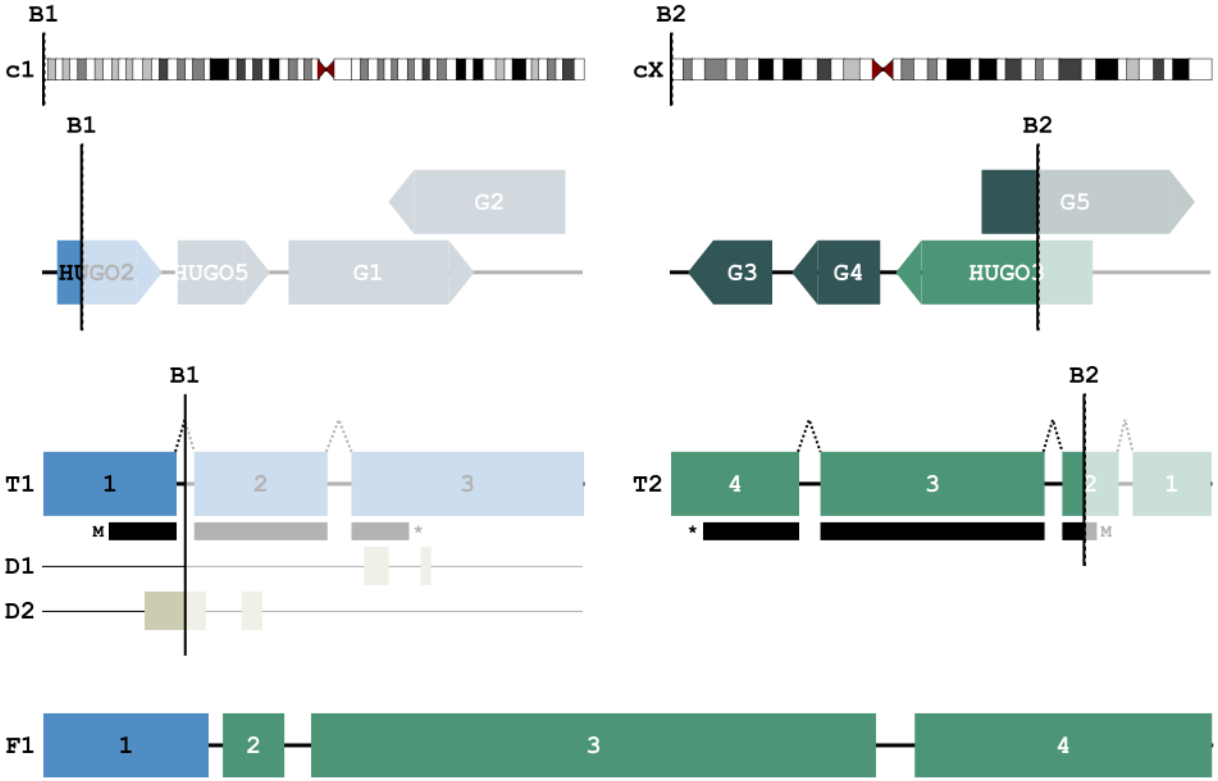
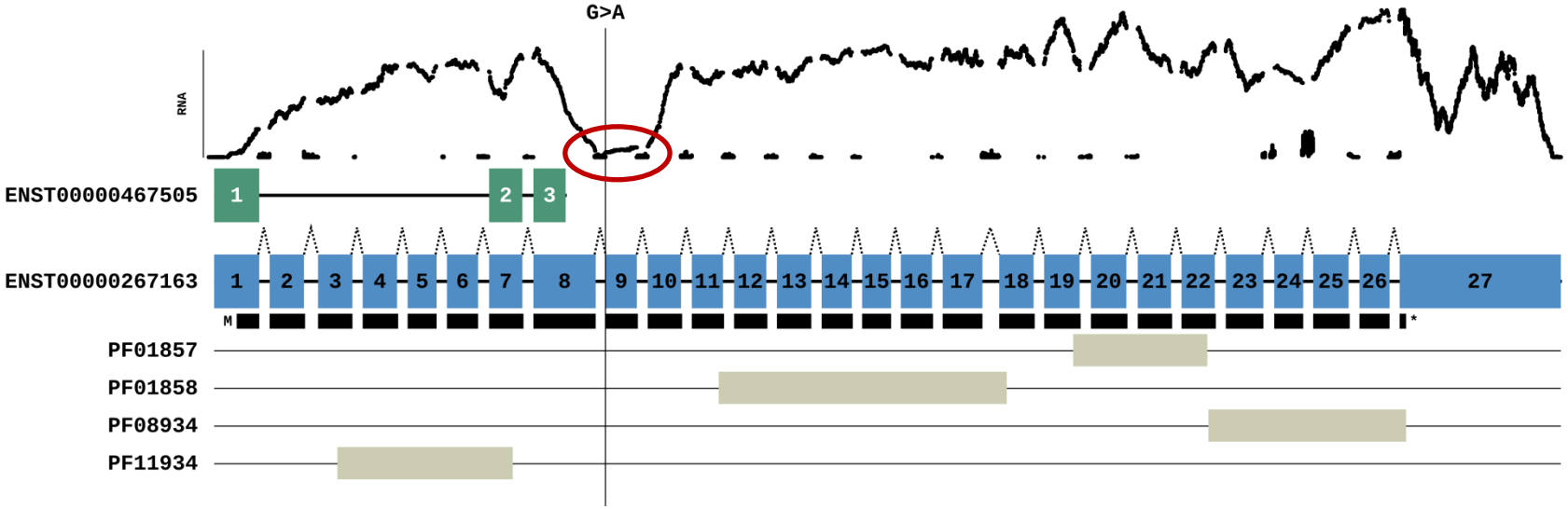


Illustration- alternative splicing



Want more information?

- MAVIS
- https://github.com/bcgsc/mavis/
- mavis@bcgsc.ca
- Submitted to Bioinformatics
- Poster presentation AGBT 2018

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

Acknowledgements

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Production Bioinformatics Analysis Groups
- Amir Zadeh
- Morgan Bye
- Simon Chan

Production Bioinformatics Database Groups
- Eric Chuah
- Tina Wong
- Dean Cheng
- Kane Tse

Yussanne Ma
Steven Jones
Marco Marra