Fusion transcript detection in rare genetic disease

Gavin Oliver
Principal Bioinformatician
Associate in Health Sciences Research
Assistant Professor of Biomedical Informatics
A working definition

• Fusion transcription involves the aberrant conjoining and expression of normally discrete genic material

• Therefore a fusion can be considered “Aberrantly conjoined and expressed genic material that exists separately under normal conditions”

• More simply: pieces of multiple genes are expressed as one

• Caused by a variety of abnormalities at the DNA level as well as (debatably) at the RNA level
Mechanisms of formation

A Translocations

Reference chromosome

Translocation

Reference genes

Fused genes
Mechanisms of formation

B Insertions

Reference chromosome

Inserted and deleted chromosome

Reference genes

Fusion genes
Mechanisms of formation

C Deletion

Reference chromosome

Deleted chromosome

Reference Genes

Fused gene

5’ 3’
Mechanisms of formation

D Tandem duplications

Reference chromosome

Aberrant chromosome

Reference genes

Fused gene

Reference genes
Mechanisms of formation

F Chromothripsy

Reference chromosome/genes

Catastrophic event (Chromosome shattering)

Rearranged chromosome/genes

Lost genomic regions
An oncogenic phenomenon?

Gene fusions in cancer
100 year timeline

- Philadelphia chromosome (CML)
- BCR-ABL1 (CML)
- t(9;22) (CML)
- t(15;17) (APL)
- t(11;22) (Ewing's sarcoma)
- PAX3-FKHR (alveolar rhabdomyosarcoma)
- AML1-X (AML)
- ATF-EWS (clear cell sarcoma)
- FUS-ERG (AML)
- PML-RARA (APL)
- CHOP-FUS (liposarcoma)
- MLL-X (MLL)
- FLI1-EWS (Ewing's sarcoma)
- ETV6-NTRK3 (Ph-like ALL)
- ERG-EWS (Ewing's sarcoma)
- SYT-SSX (synovial sarcoma)
- NPM-ALK (anaplastic lymphoma)
- PDGFb-TEL (CML)

Hematological
Soft tissue
An oncogenic phenomenon?

Gene fusions in cancer
100 year timeline

- CTNNB1-PLAG1 (salivary)
- HMGA2-NFIB (salivary)
- BRD4-NUT (midline)
- CRTC-MAML2 (salivary)
- TMPRSS2-ERG (prostate)
- SLC34A2-ROS EML4-ALK (lung)
- TMPRSS2-ETV1
- TMPRSS2-ETV4
- TMPRSS2-ETV5 (prostate)
- X-MASTx
- X-NOCTHx (breast)
- CD44-SLC1A2 (gastric)
- X-FGFRx (cholangio, breast, lung, prostate, thyroid, oral, bladder)
- FGFRx-X
- MYB-NFIB (ACC)
- LACTB2-NCOA2 (CRC)
- CLDN18-ARHGAP (gastric)
- X-SKIL (prostate)
- X-RSPO2 (prostate)
- X-PK3CA/B (prostate)

X- Multiple partner genes
x- Multiple isoforms of the same gene

Rare epithelial

Common epithelial
Mechanism of action

• Commonly
  – involve fusion of a downstream kinase
  – or transcription factor
    • with a more highly expressed upstream gene
      – leading to increased expression of the downstream gene or a functional component of it

• Protein formation dependent on in-frame translation
Mechanism of action

• Most reported gene fusions pertain to gain-of-function aberrations imparting neoplastic phenotypes

• Loss of function of tumor suppressors such as TP53 and PTEN have also been identified

• Fusion transcripts are recognized as having diagnostic, prognostic and therapeutic (druggable) relevance in oncology

• Detection of gene fusions is increasingly incorporated into the standard workflow for genomic characterization of tumors in both research and clinical settings
Fusions in inherited disease

- 18-40% unsolved cases are solved by exome sequencing
- RNA-Seq has recently been proposed as a supplementary diagnostic tool
- Cummings et al. achieved a 35% diagnostic increase by profiling aberrant splicing and allele specific expression
- Kremer et al. added gene expression quantification to the testing repertoire and demonstrated a 10% increase
- Isolated reports exist in the literature of fusion transcripts being detected in cases of brain malformation, intellectual disability, schizophrenia, ASD and more
- Fusion transcription had not been systematically profiled in inherited disease
Patient Cohort

- 47 patients
- Prior exome-sequencing
- 23 M, 24F
- Ages 9 months – 68 years (median 11)
- Diverse phenotypes
  - Neurological
  - Muscular
  - Gastrointestinal
  - Skeletal
  - Connective tissue disorders

Oliver et al. (2019) https://doi.org/10.1371/journal.pone.0223337
RNA-Seq in inherited disease

RNA-Seq in inherited disease

- Patient whole blood
- Illumina HiSeq 2500
- 200 million 100bp PE reads per sample
Fusions in inherited disease

• Fusion detection increased diagnosis of rare disease
  – Two cases confirmed solved
    • SCID
    • Multiple exostoses
  – 4.3% increase in diagnostic yield
  – Experimentally validated existence of fusion events in disease-relevant genes with potential phenotypic relevance in five additional cases
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Software solution overview
Fusion transcripts

Gene A (Chromosome 1)

Gene B (Chromosome 21)
Fusion transcripts

Gene A (Chromosome 1)

Gene B (Chromosome 21)
Fusion transcripts

Gene A (Chromosome 1)

1 2 3 4 5 6 7

Gene B (Chromosome 21)

1 2 3 4

Fusion transcripts
Fusion calling challenges

• Complicated by the many false positive candidates resulting from:
  – alignment artifacts such as multi-mapping of reads owing to homologous (pseudogenes) and/or repetitive sequences
  – sequencing artifacts due to errors in library generation (particularly ligation and PCR artifacts) and sequencing

• Incorporating these considerations, and additional bioinformatics filters, various bioinformatics pipelines have been developed to help prioritize fusion candidates from next-generation sequencing (NGS) data

• “Read-through” transcription of neighboring genes occurs frequently in normal cells

• Common non-pathogenic fusion events between distal genes are known to exist due to distinct polymorphic haplotypes
Fusion calling challenges

• Numerous software solutions exist for fusion detection
  – e.g. STAR-Fusion, Tophat-Fusion, PRADA, Fusioncatcher

• Technical comparisons demonstrate limited overlap and no caller is fully inclusive
  – Partially because FPs are abundant & outputs require filtering
  – Filters are trained using *in-silico*, tumor or cell-line data & performance falters on alternative data types

• It is recommended to select a caller on the basis of the data being profiled however none are trained on inherited disease
Fusion calling challenges

- Any attempt to detect fusions in inherited disease thus requires:
  - Inherent sensitivity
  - A means of deprioritizing biologically and phenotypically unimportant fusion candidates
Filtering / Prioritization

Read support (basic)

- **TopHat Fusion** *(Kim & Salzberg 2011)*
  - Equally applicable to other callers

- Omitted all TopHat filtering steps (cancer cell-line derived)

- Employed a very minimal depth filter (2 reads)
Normal DB comparison

- Compared fusion candidates to a database of candidates from normal tissues
- Fusion calling on samples from GTEx, Illumina Human Bodymap, Mayo Clinic
- Approx. 800 samples, 30 tissues
- Any fusion candidates occurring in DB or more than one cohort sample were categorized as normal/recurrent
Phenotypic Prioritization

- Dual approach
  - Manual (Literature, OMIM, Genecards)
  - *In-silico*
    - **PCAN:** phenotype consensus analysis to support disease-gene association (Godard & Page, 2016)

- Generated phenotypically prioritized events for follow-up validation
BLAST categorization

- Fusion consensus sequences generated by TopHat Fusion used as input
  - Algorithm dependent

- Devised custom categorization pipeline based on BLASTn

- Categorization logic based on best alignments
Candidate Categorization

- Fusion consensus sequence
- BLASTn
- Human genome
- Human transcriptome

Potential novel transcript or intragenic fusion

Blood abundant products

Likely artefactual Potential immune diversity or homology artefact
Now let’s try it…

Questions