MAYO CLINIC

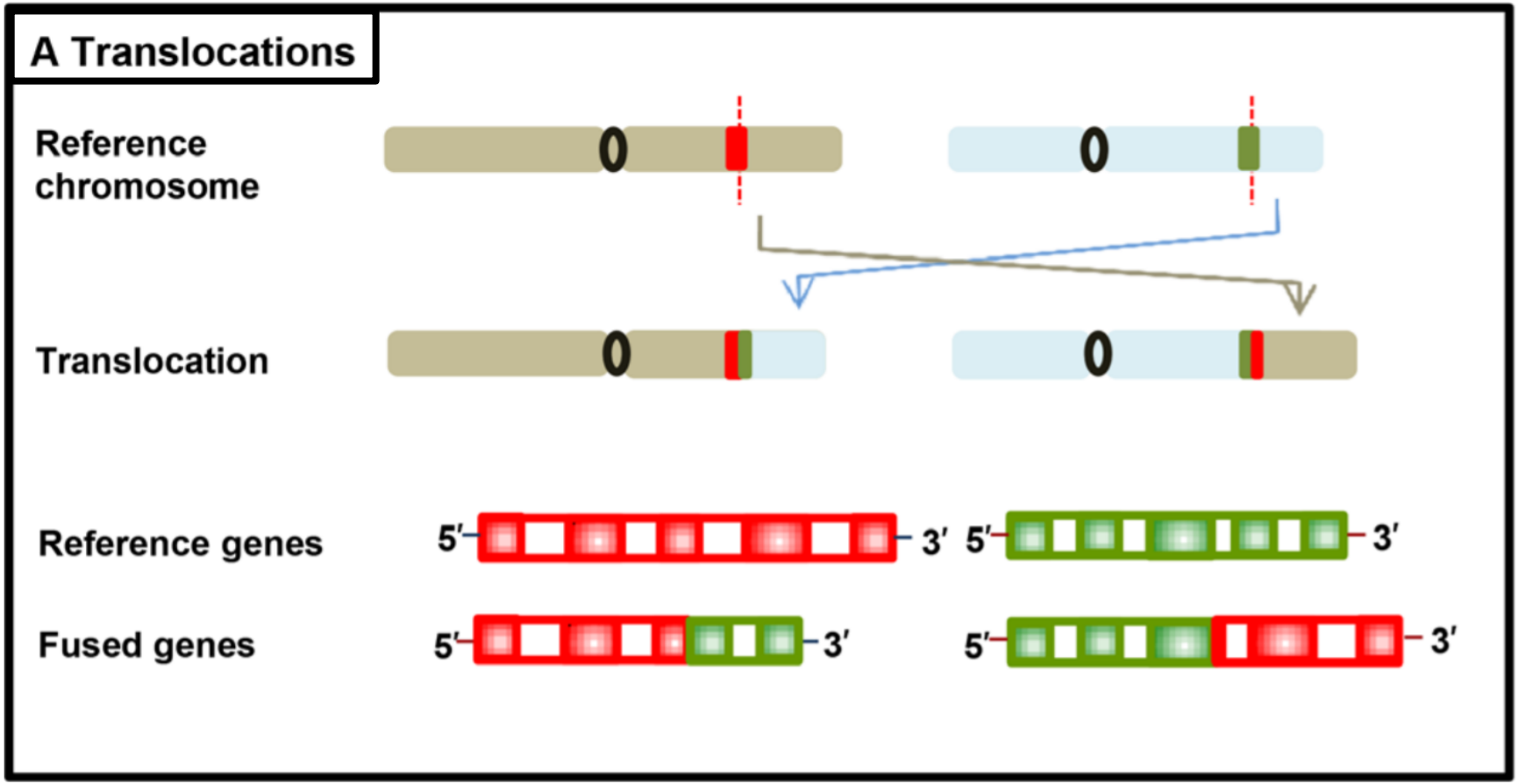
Fusion transcript detection in rare genetic disease

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

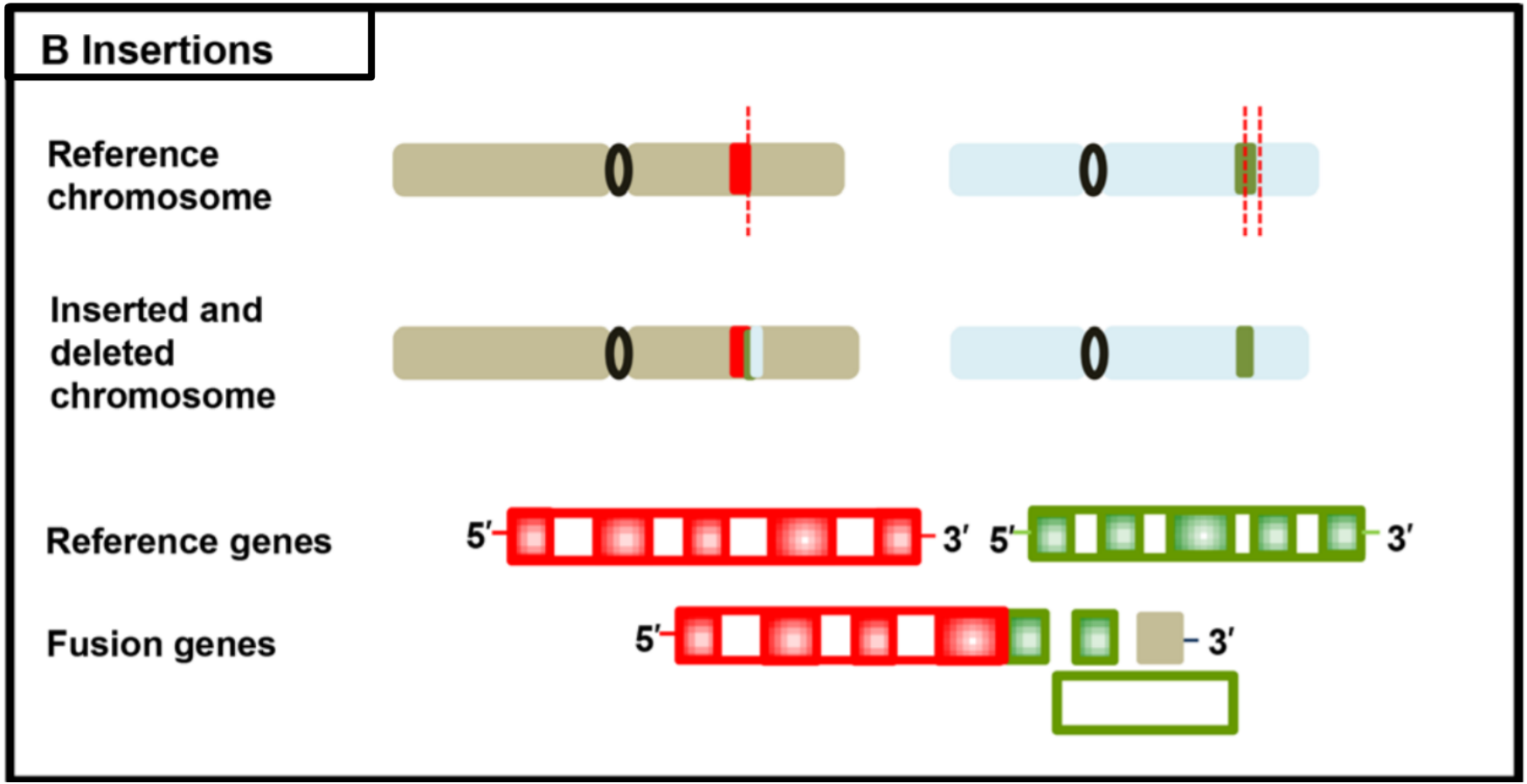
A working definition

- 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

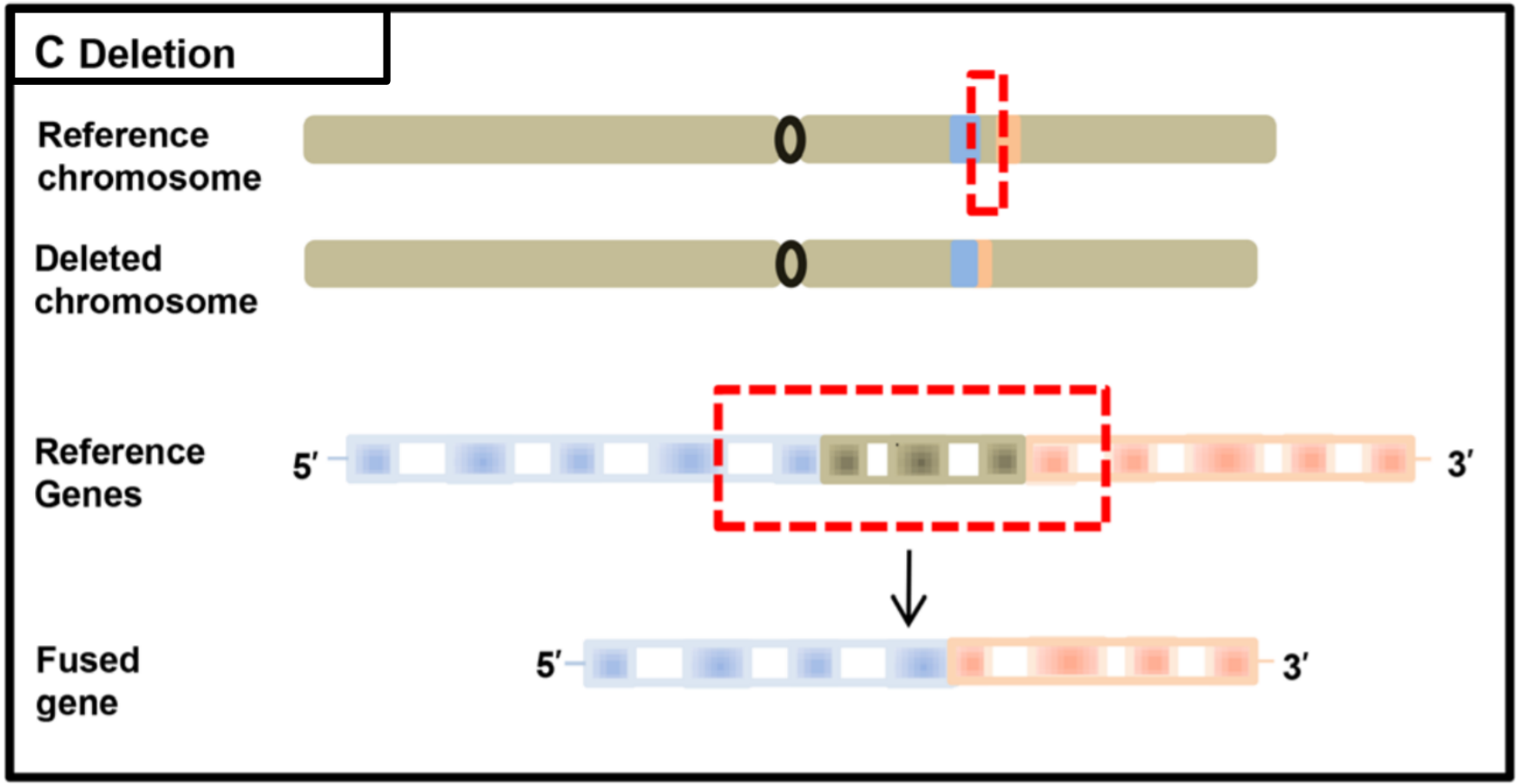
Mechanisms of formation



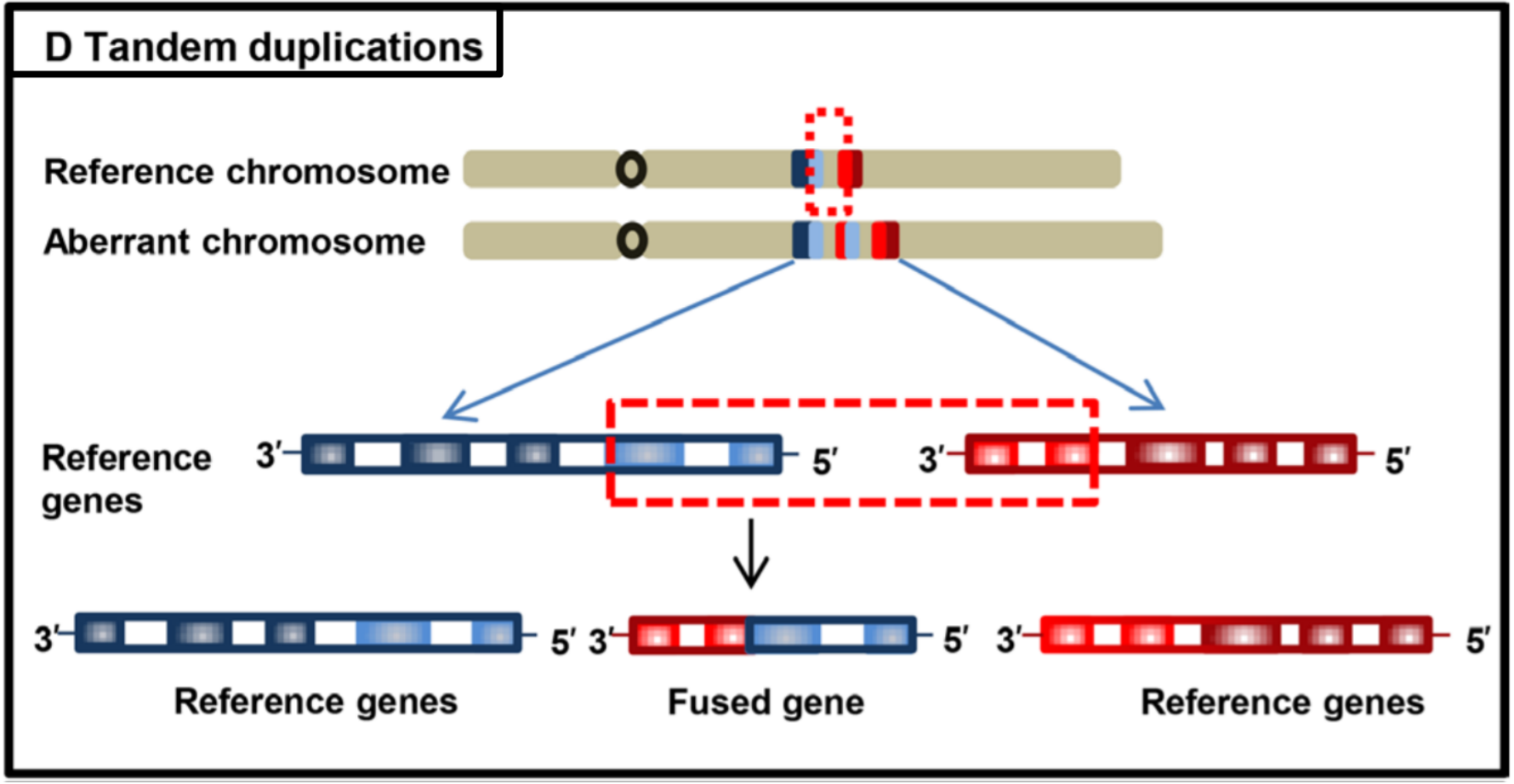
Mechanisms of formation



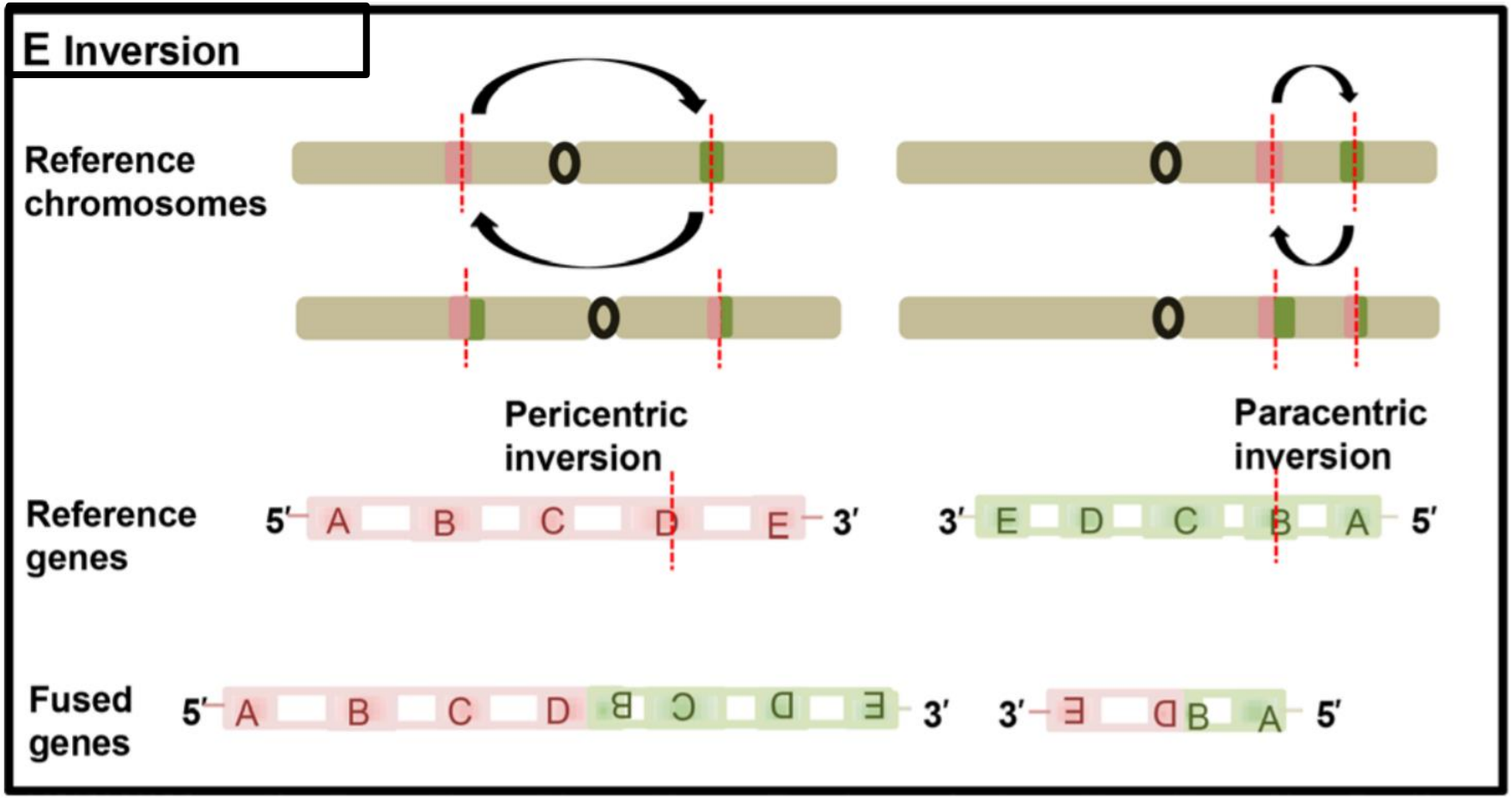
Mechanisms of formation



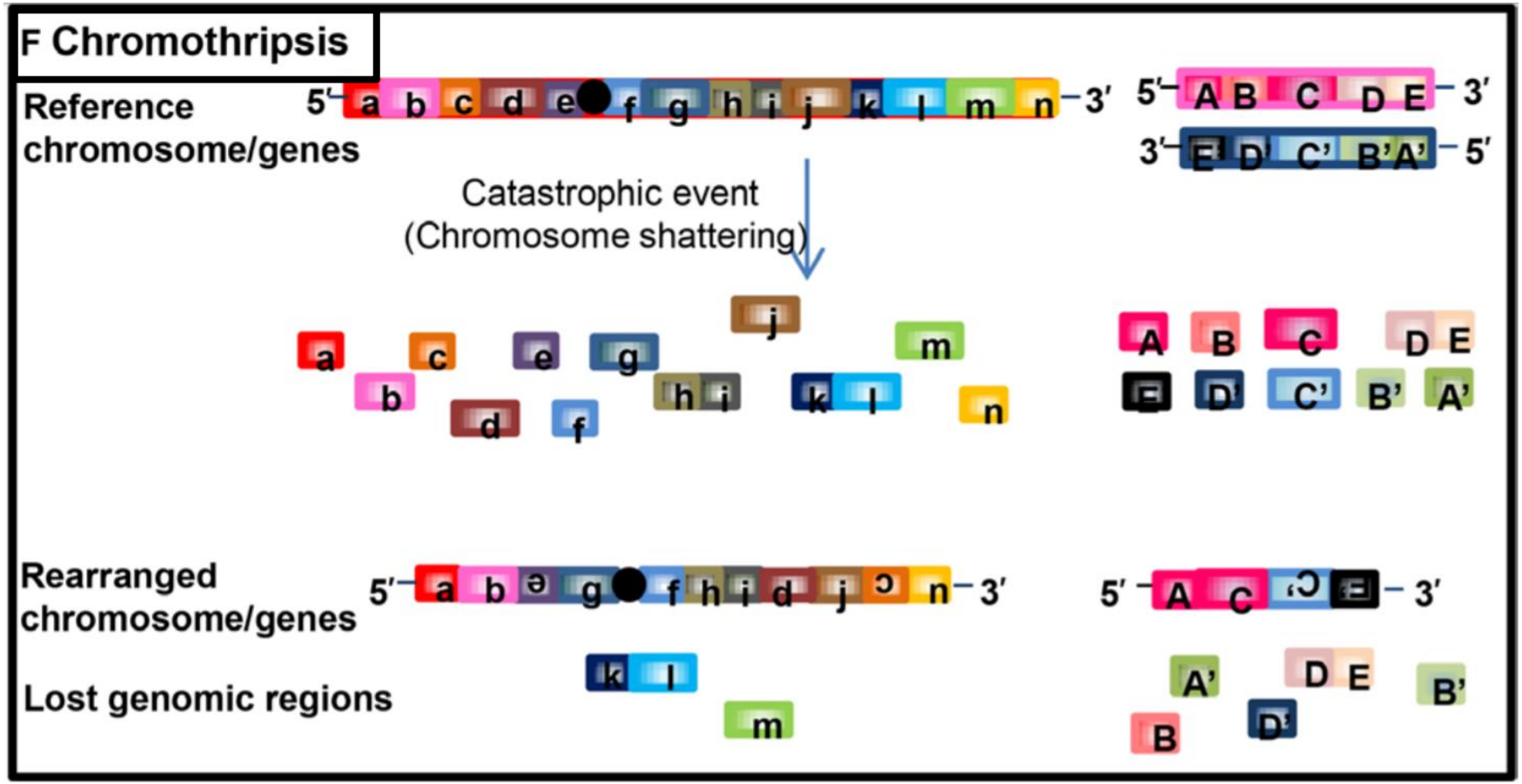
Mechanisms of formation



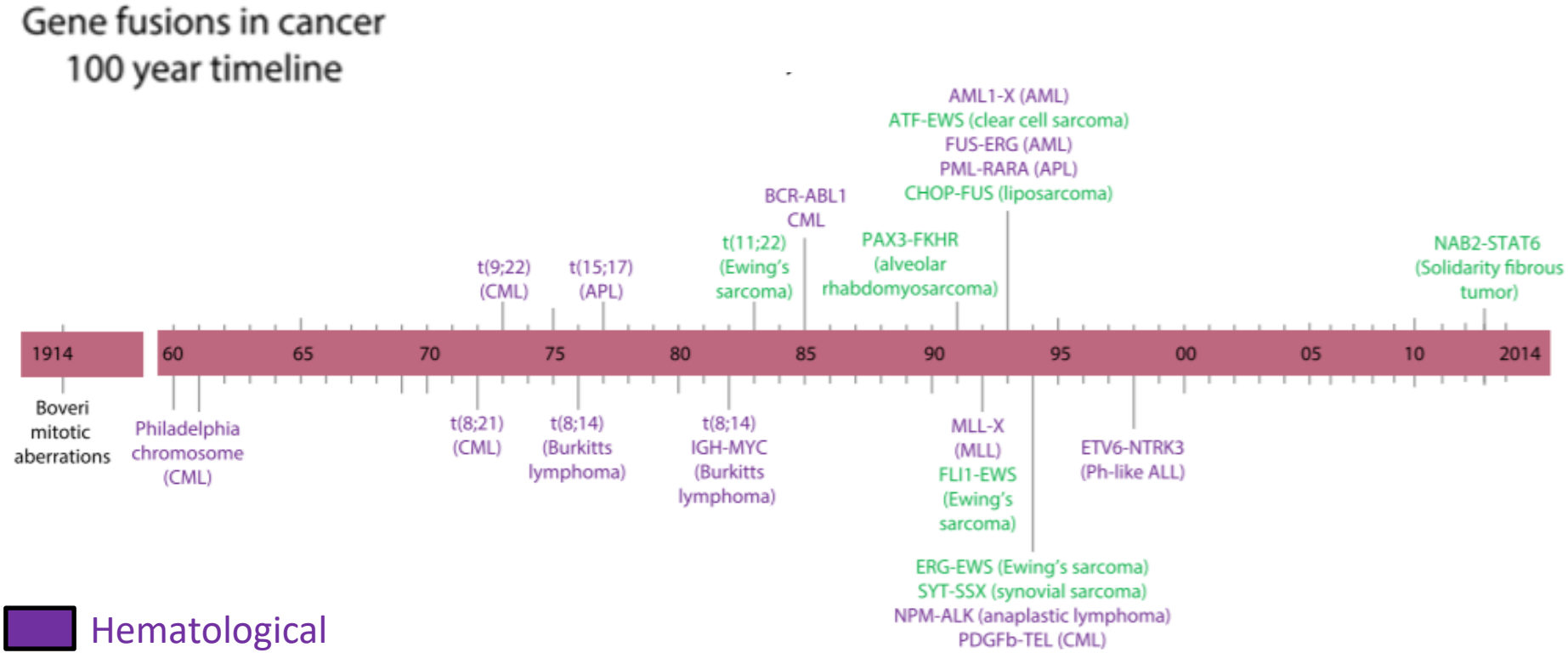
Mechanisms of formation



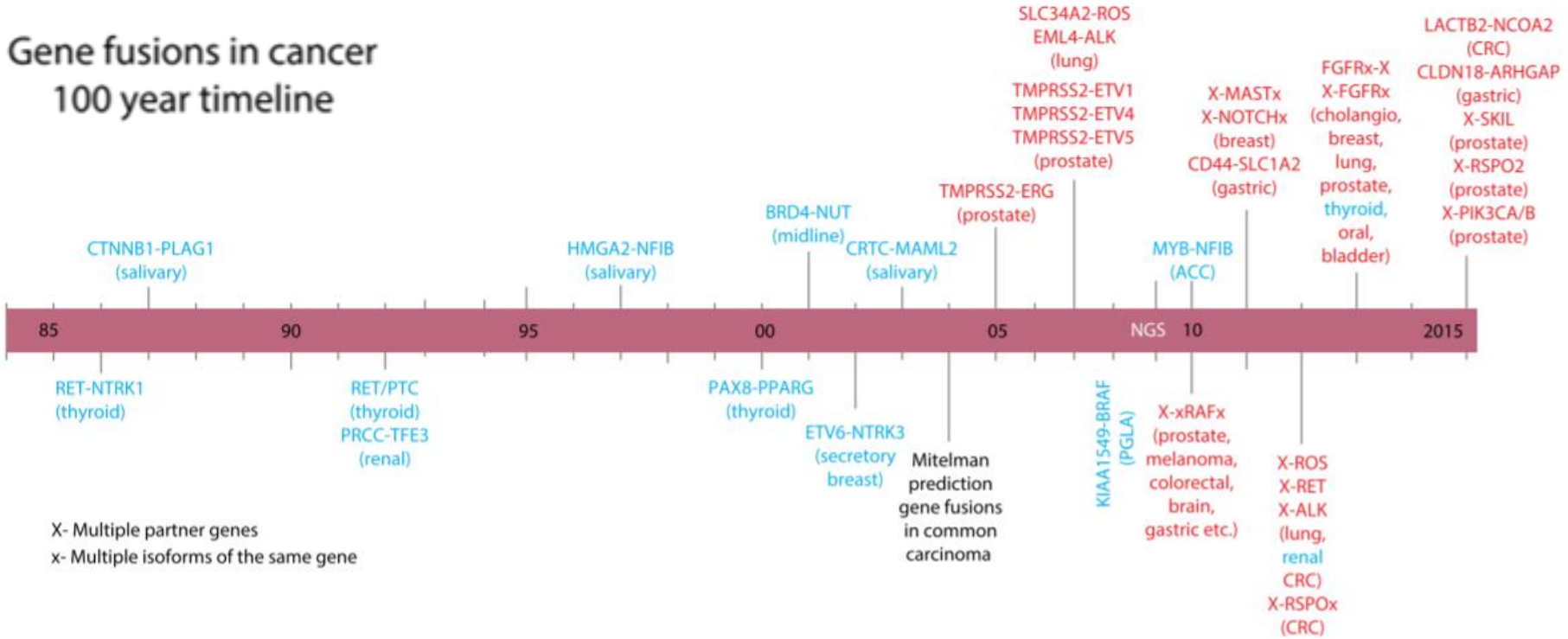
Mechanisms of formation



An oncogenic phenomenon?



An oncogenic phenomenon?



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

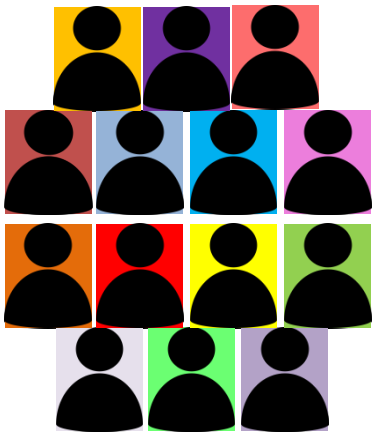
- Most reported gene fusions pertain to gain-of function aberrations imparting neoplastic phenotypes
- Loss of function of tumor suppressors such as TP53 and PTEN have also been identified
- Fusion transcripts are recognized as having diagnostic, prognostic and therapeutic (druggable) relevance in oncology
- 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

RNA-Seq in inherited disease

Patient Cohort

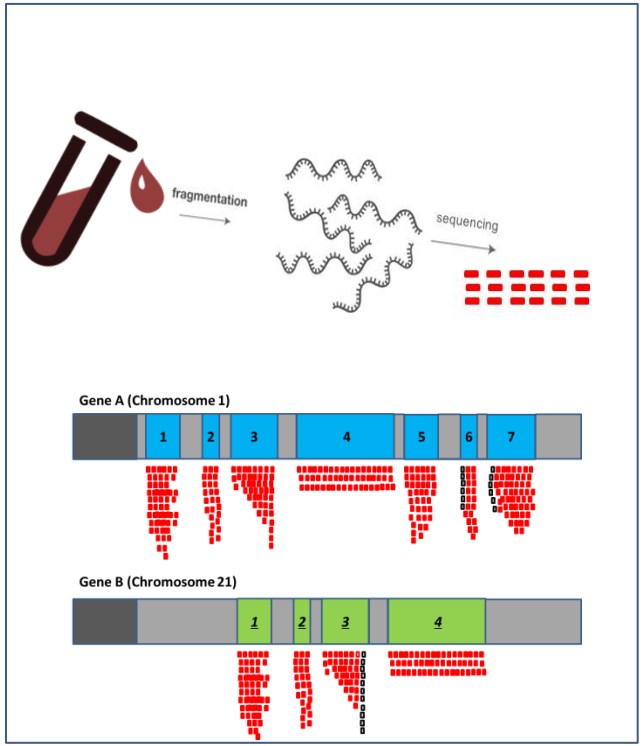


- 47 patients
- Prior exome-sequencing
- 23 M, 24F
- Ages 9 months – 68 years (median 11)
- Diverse phenotypes
  - Neurological
  - Muscular
  - Gastrointestinal
  - Skeletal
  - Connective tissue disorders

Oliver *et al.* (2019) https://doi.org/10.1371/journal.pone.0223337

RNA-Seq in inherited disease

RNA-Sequencing

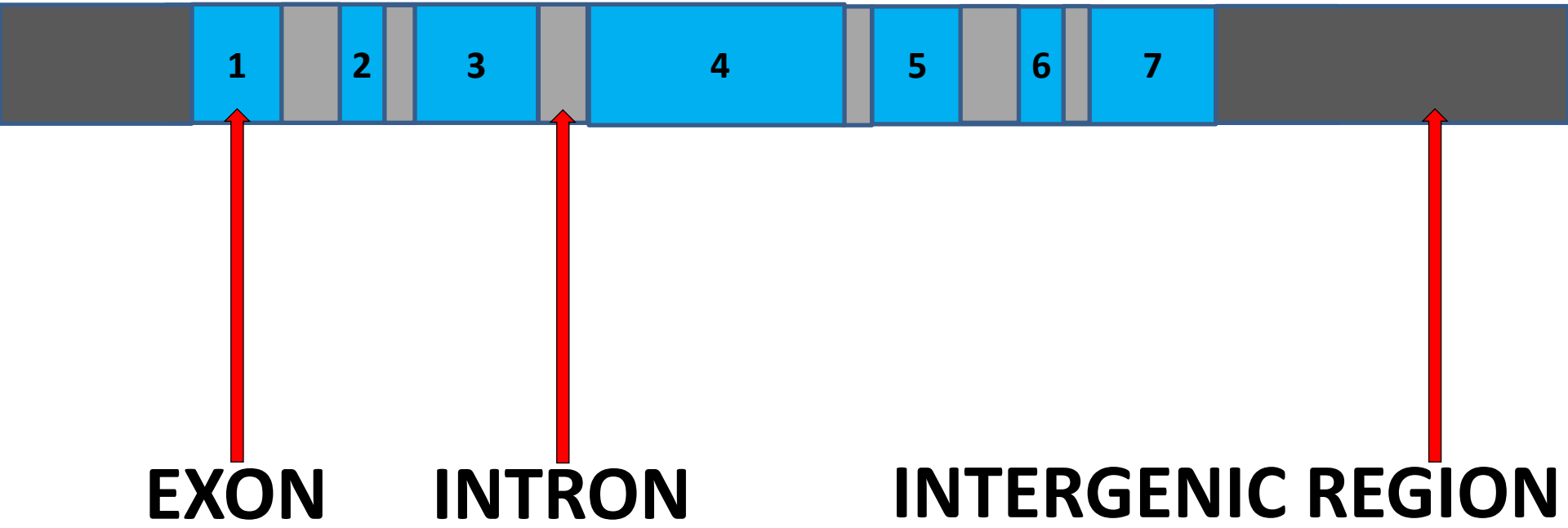


- Patient whole blood
- Illumina HiSeq 2500
- 200 million 100bp PE reads per sample

Fusions in inherited disease

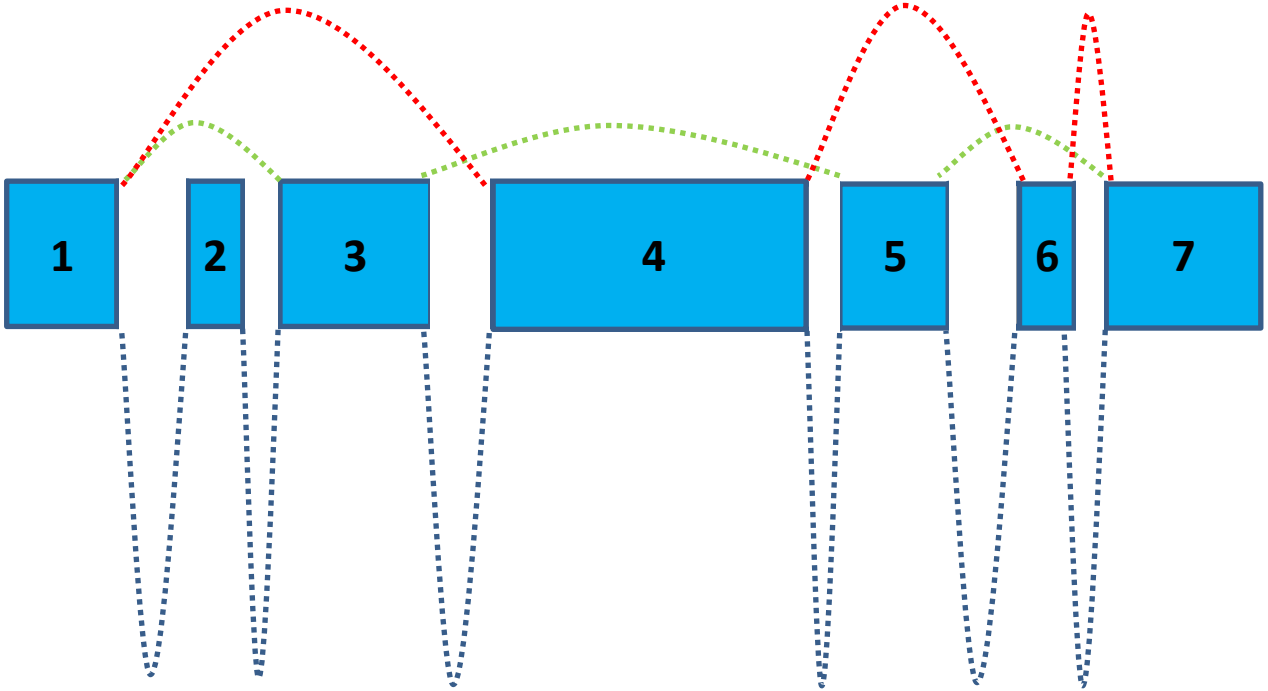
- Fusion detection increased diagnosis of rare disease
  - Two cases confirmed solved
    - SCID
    - Multiple exostoses
  - 4.3% increase in diagnostic yield
  - Experimentally validated existence of fusion events in disease-relevant genes with potential phenotypic relevance in five additional cases

Software solution overview

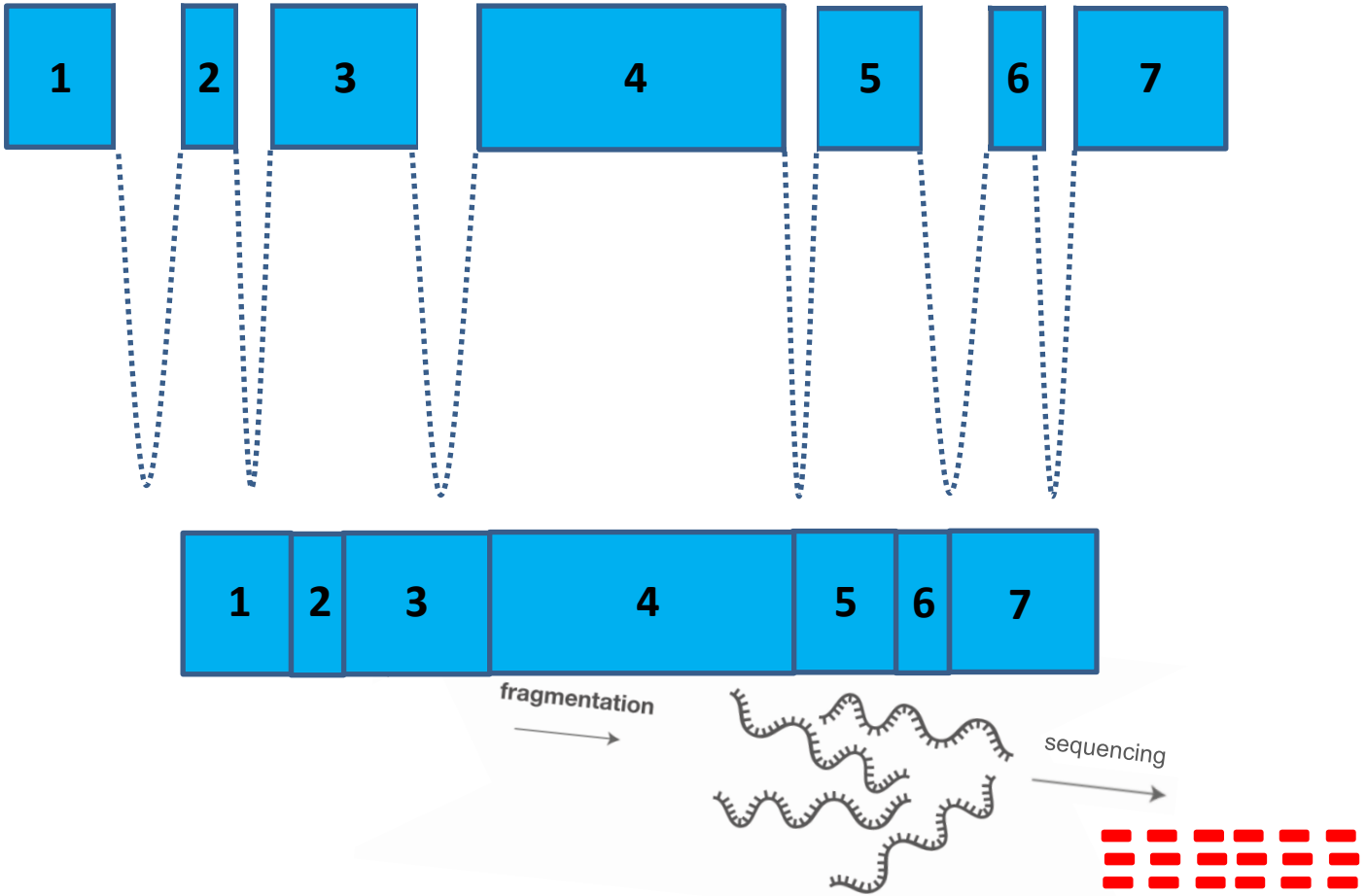


Software solution overview

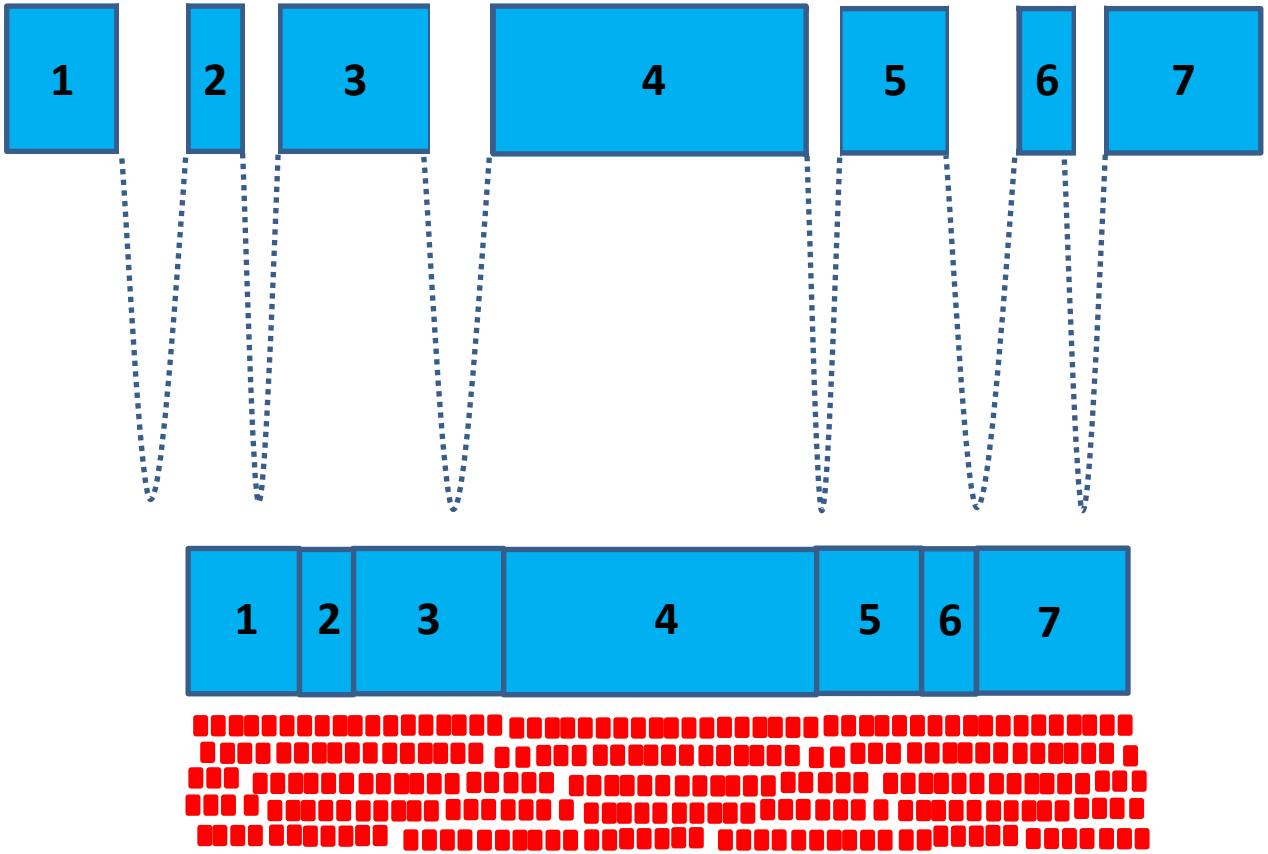
Software solution overview



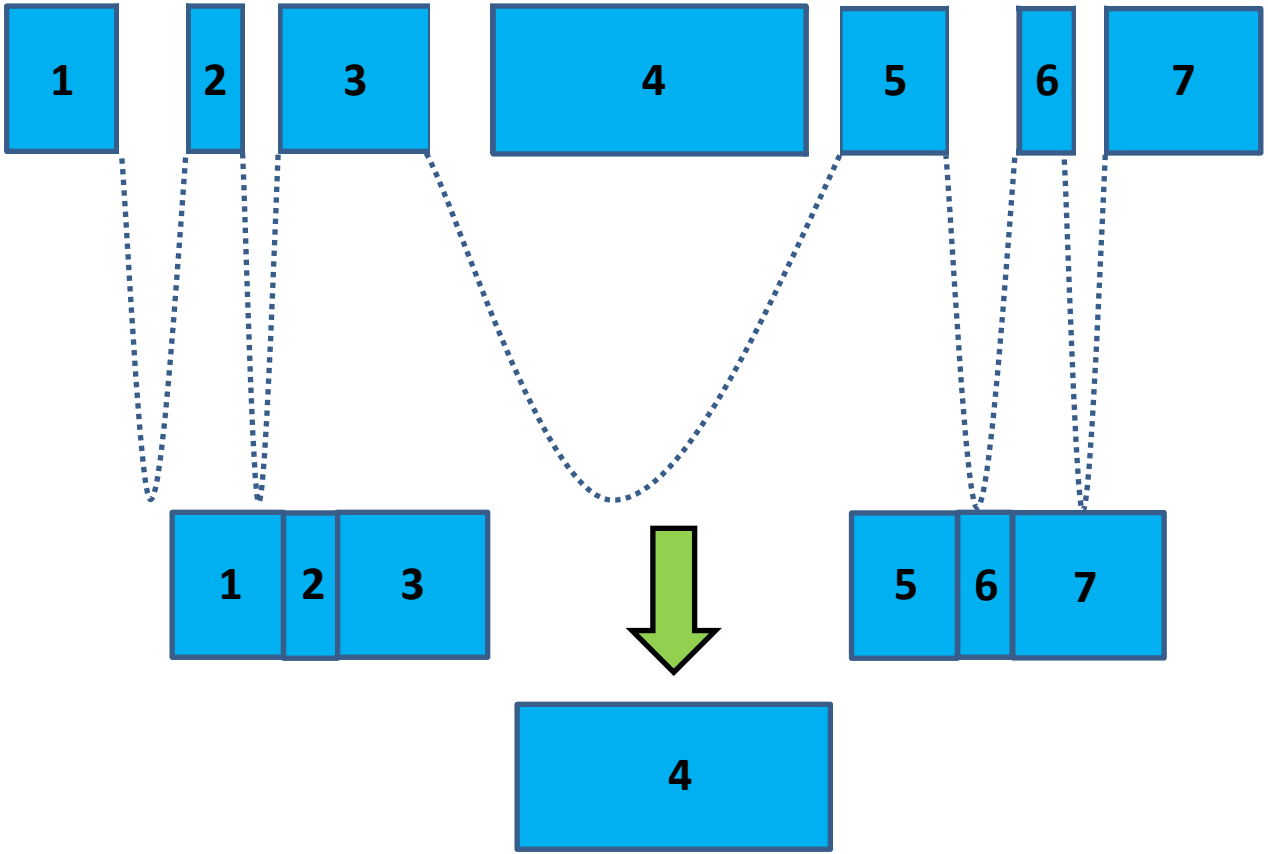
Software solution overview



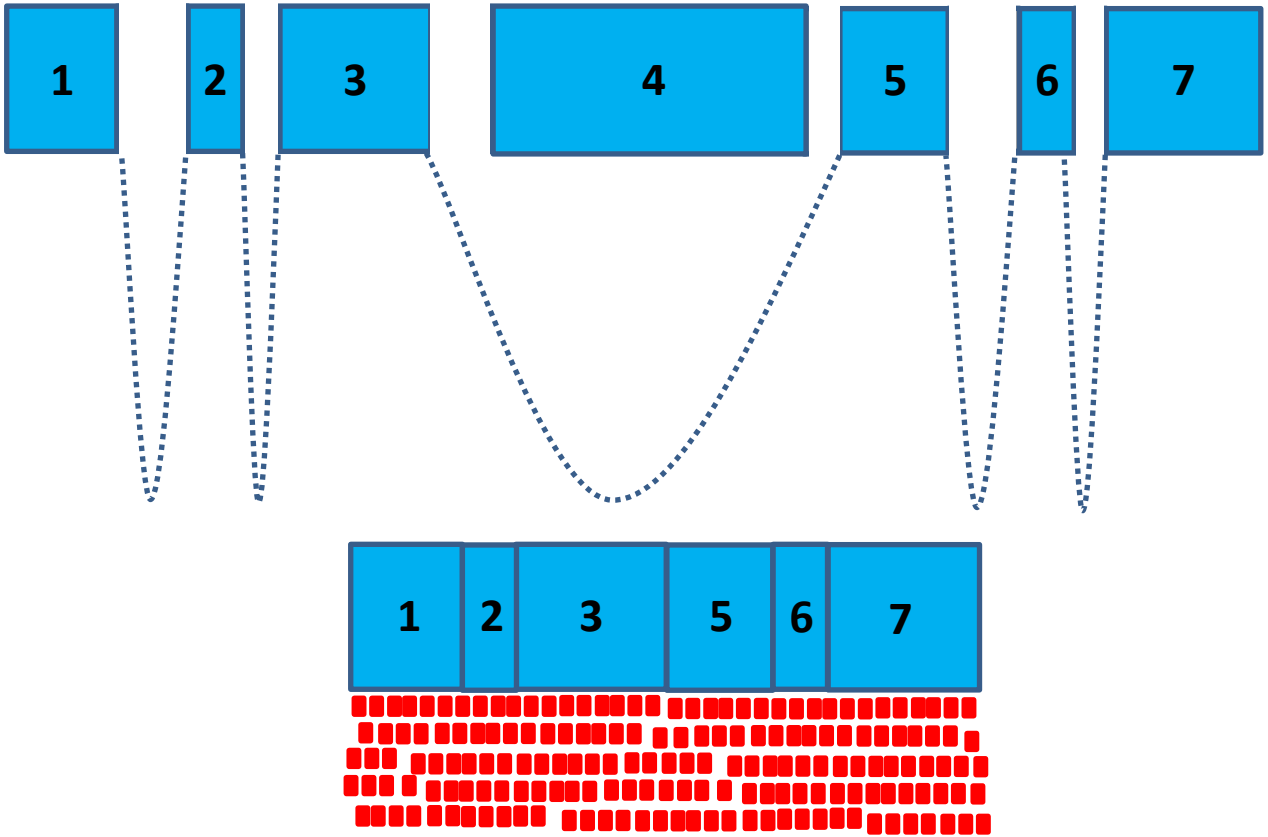
Software solution overview



Software solution overview

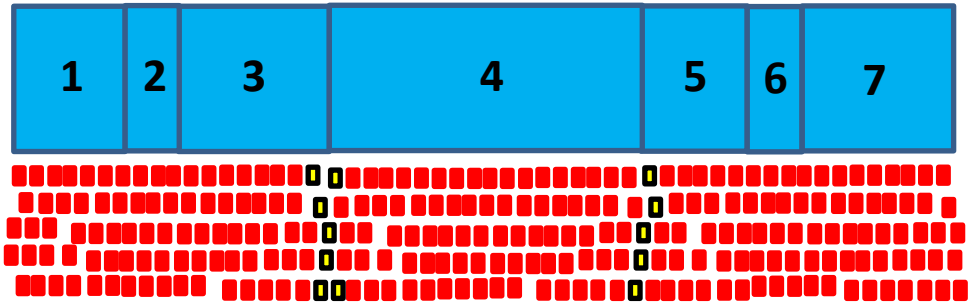


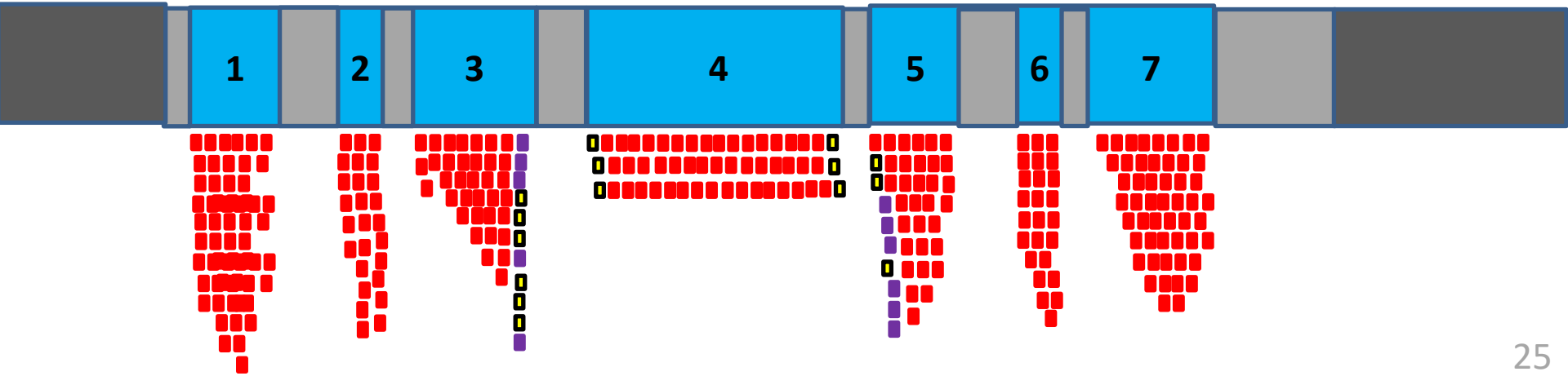
Software solution overview



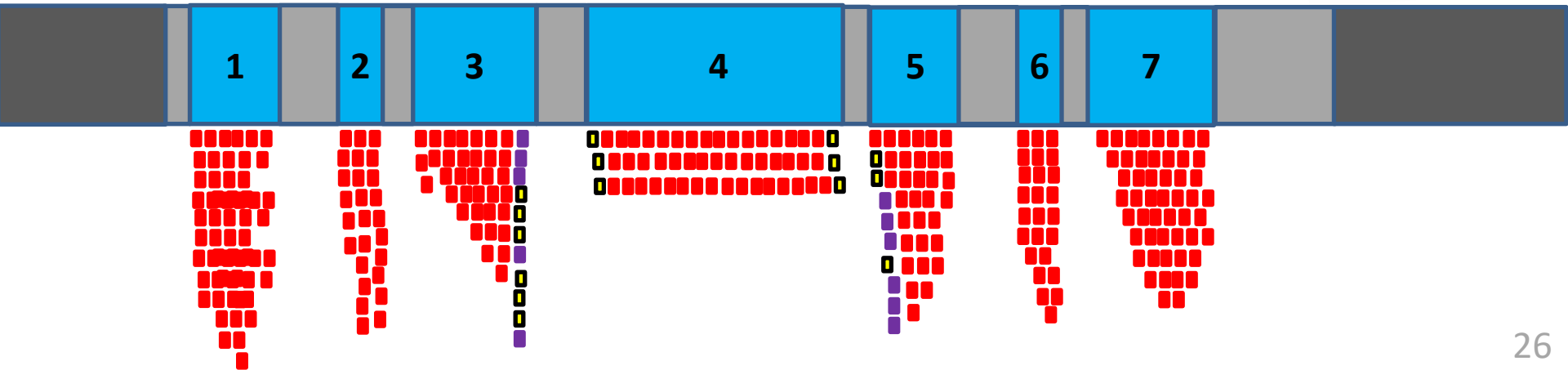
Software solution overview

Software solution overview

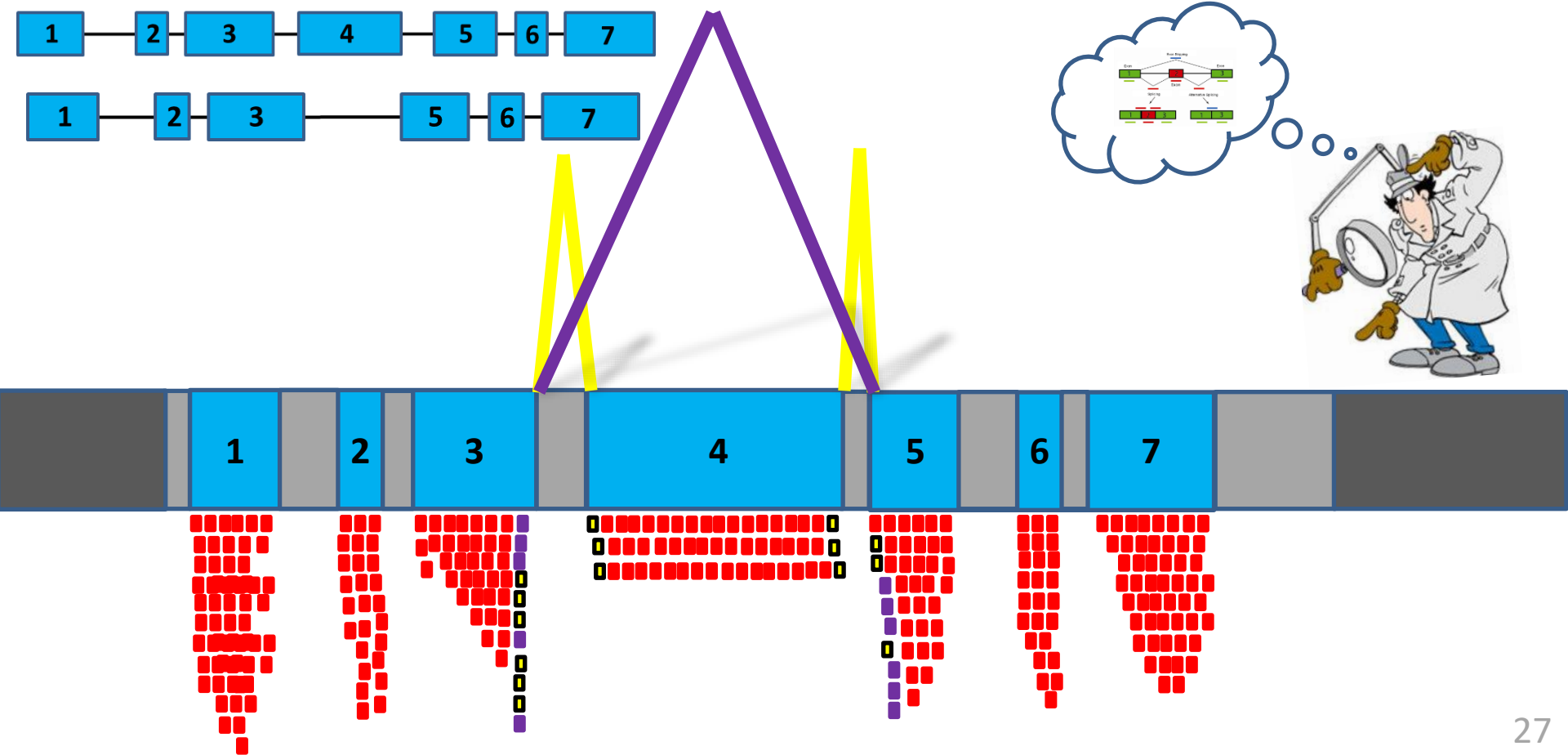




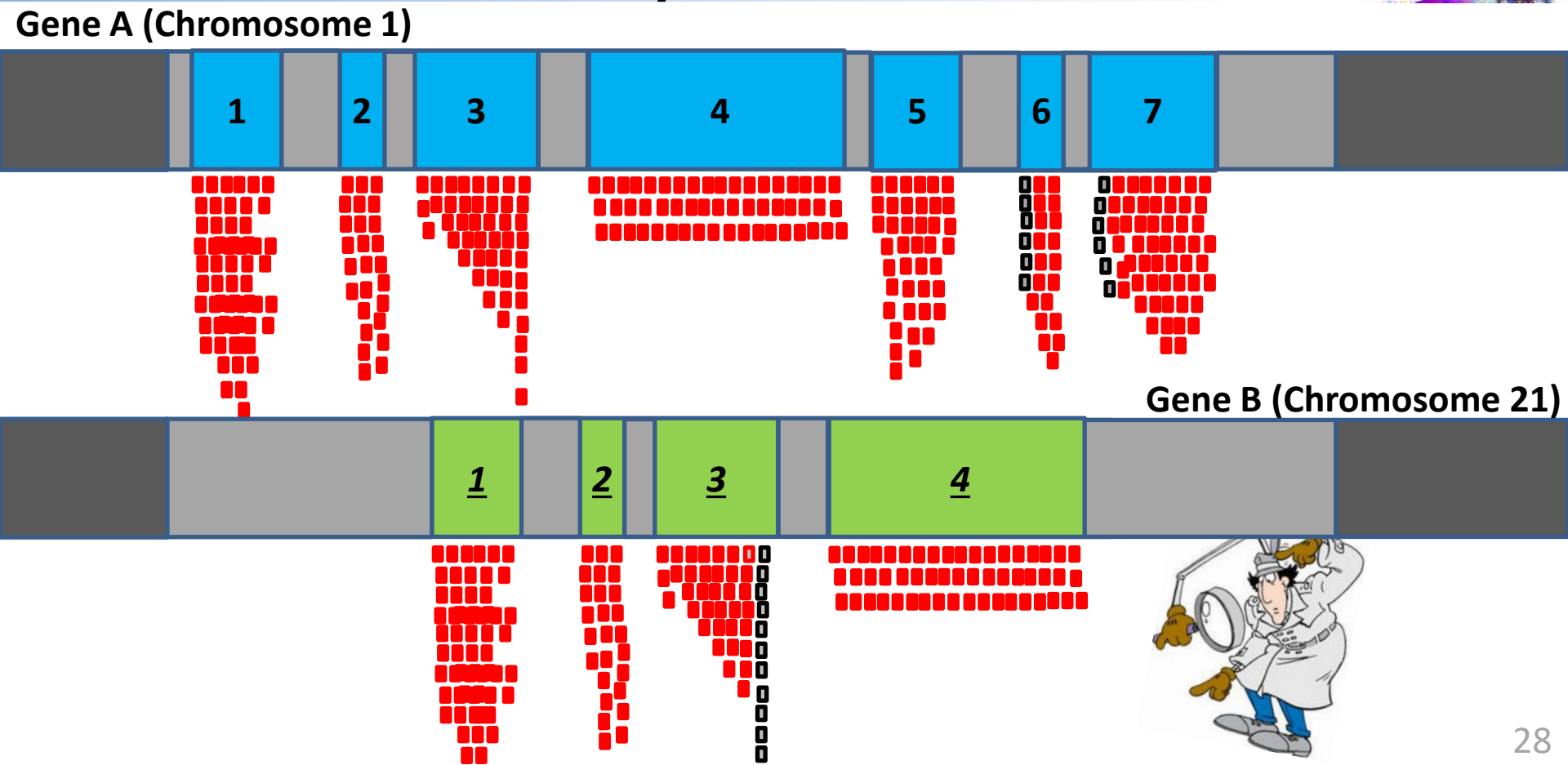
Software solution overview



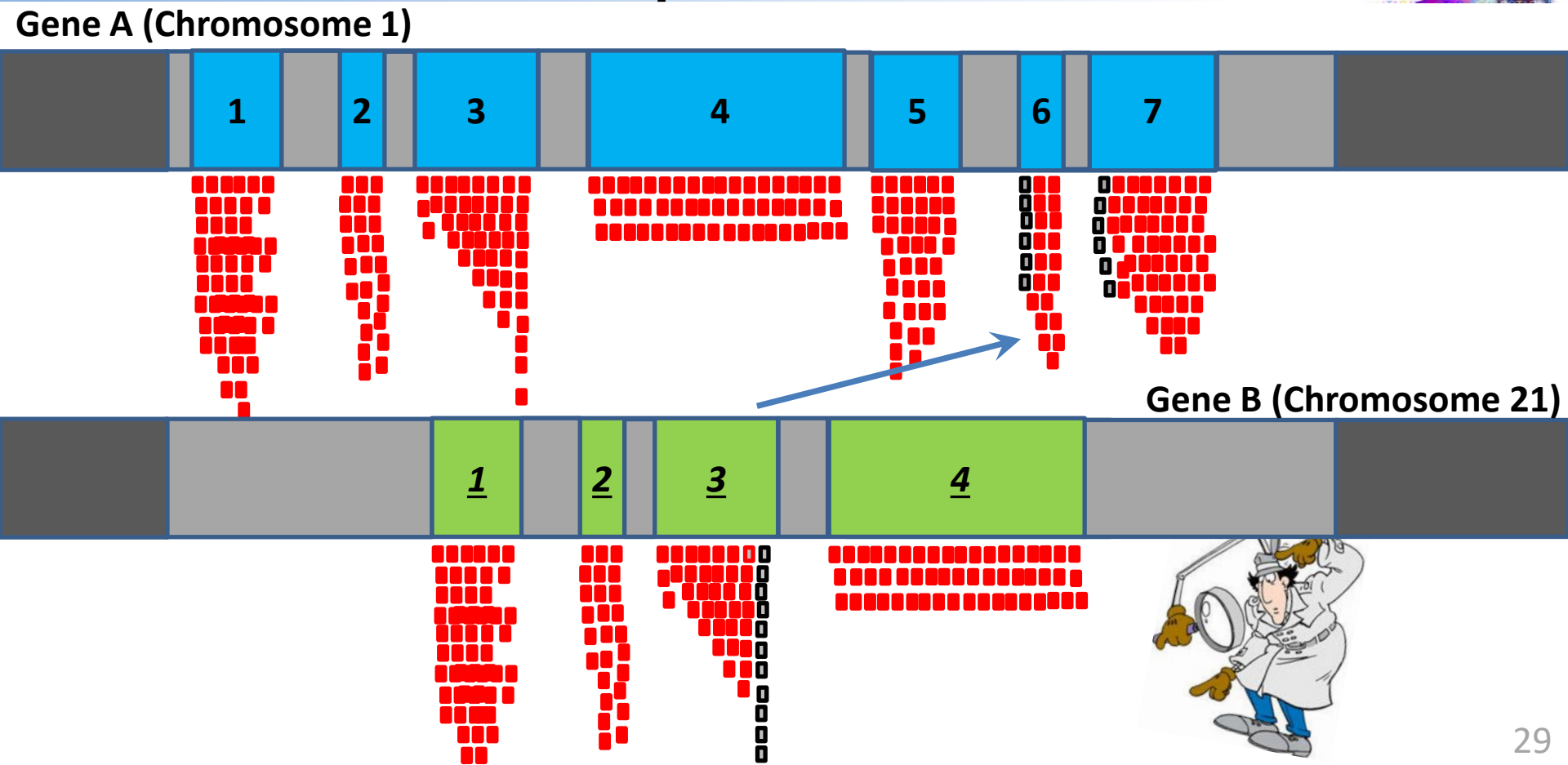
Software solution overview



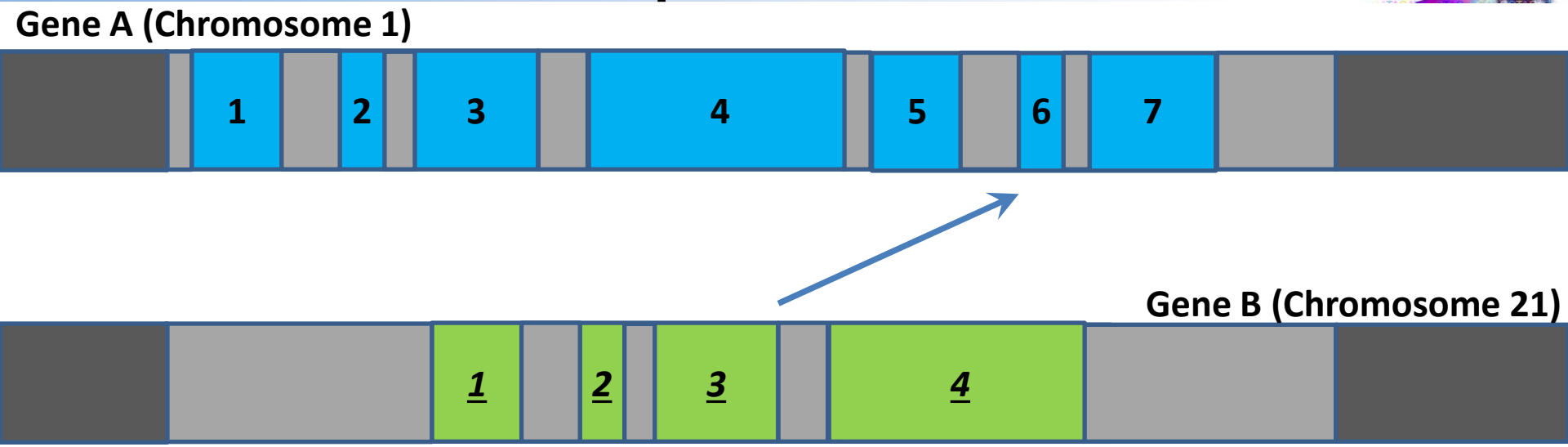
Fusion transcripts



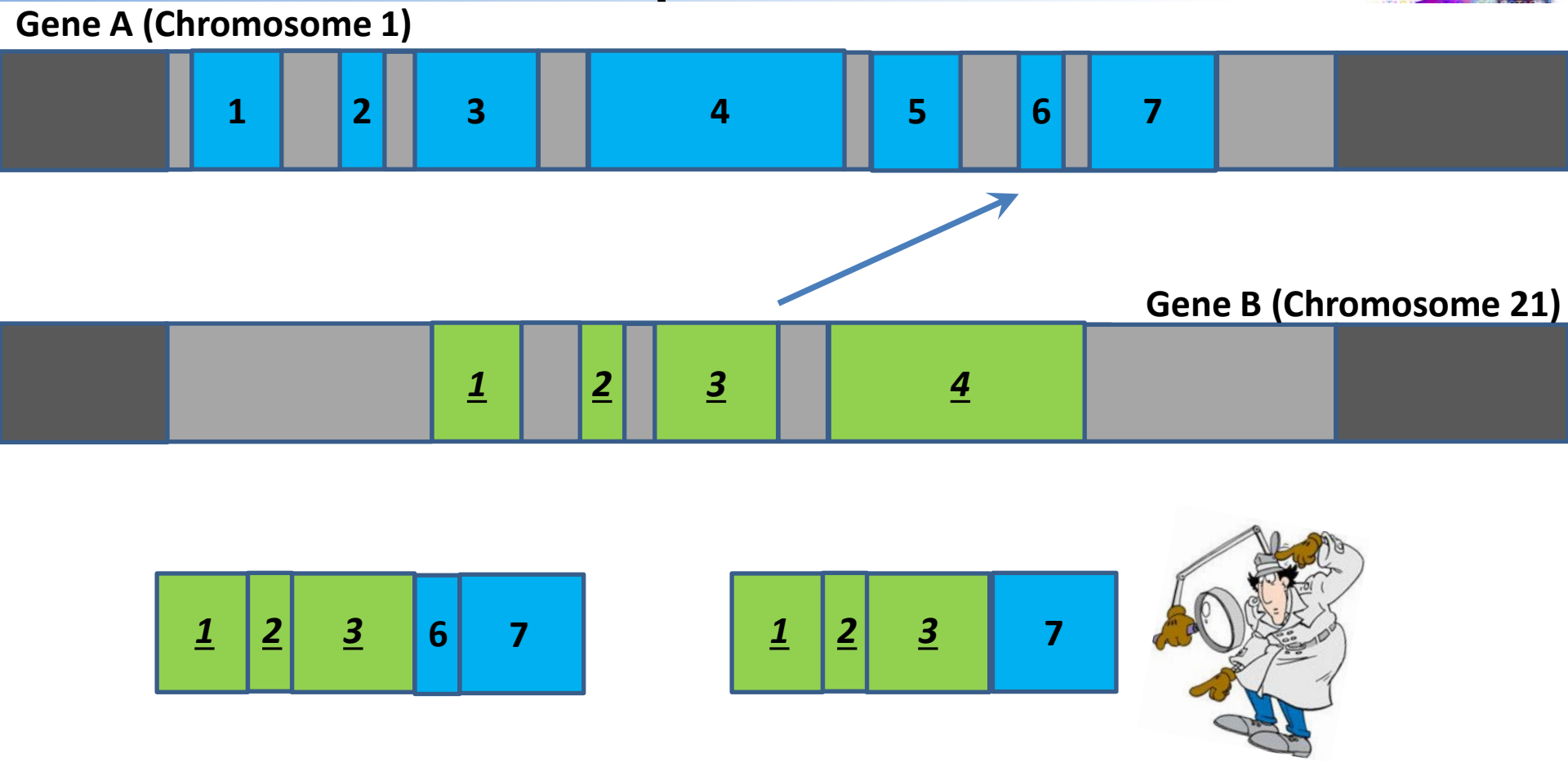
Fusion transcripts



Fusion transcripts



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

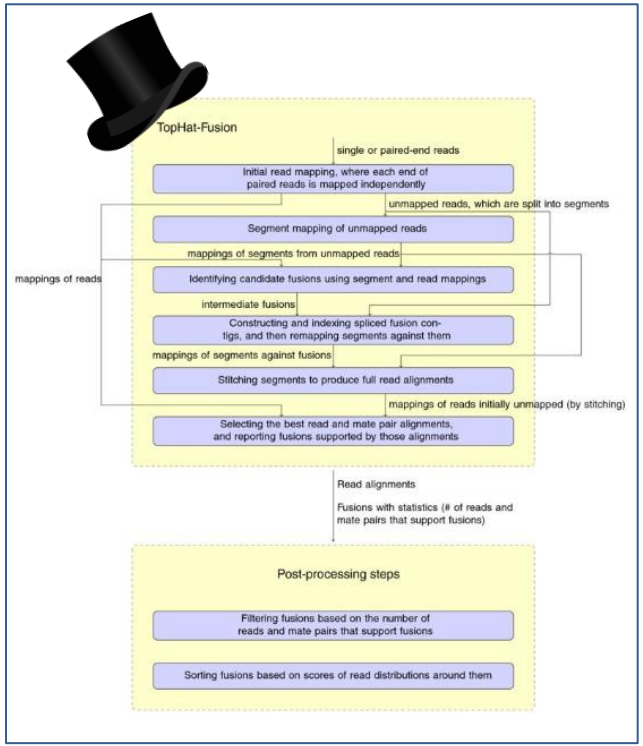
- It is recommended to select a caller on the basis of the data being profiled however none are trained on inherited disease

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

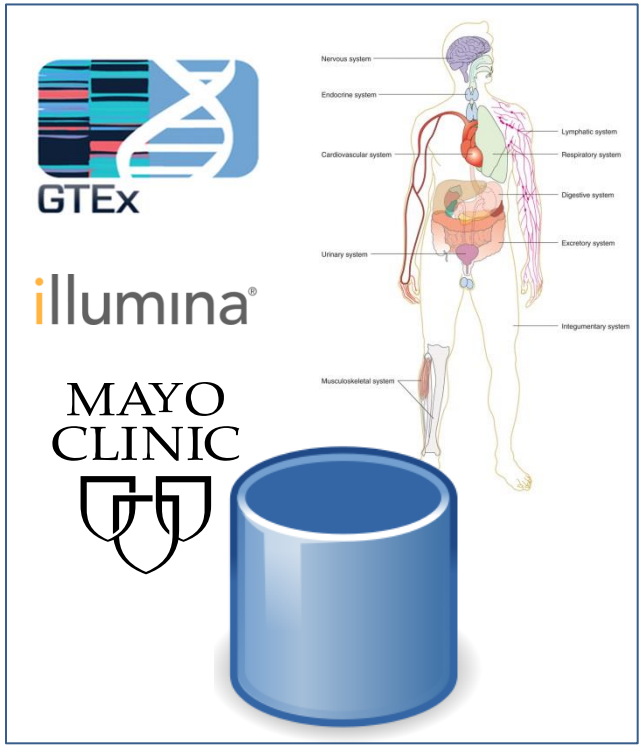
Read support (basic)



- 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)

Filtering / Prioritization

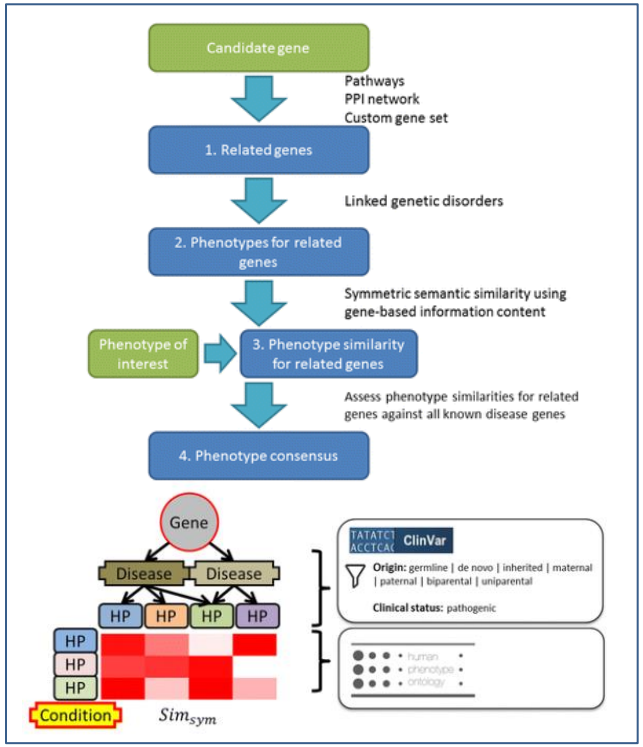
Normal DB comparison



- 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

Filtering / Prioritization

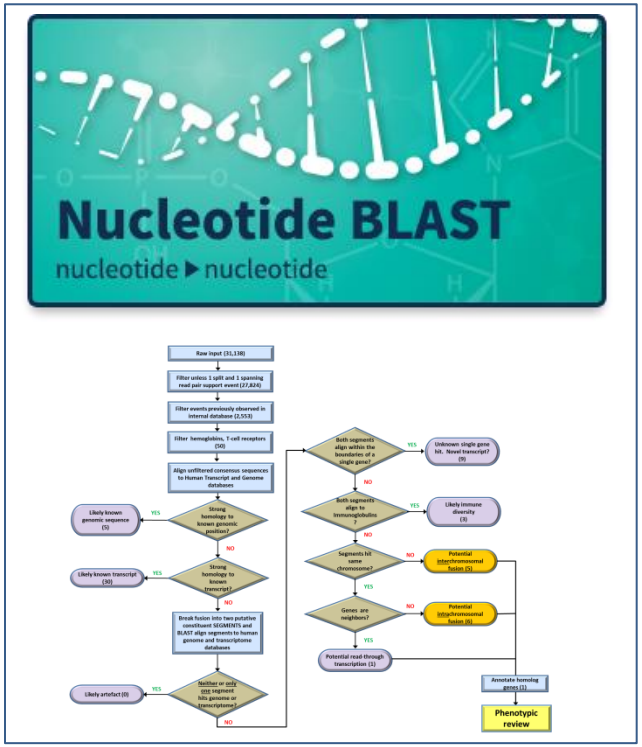
Phenotypic Prioritization



- 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

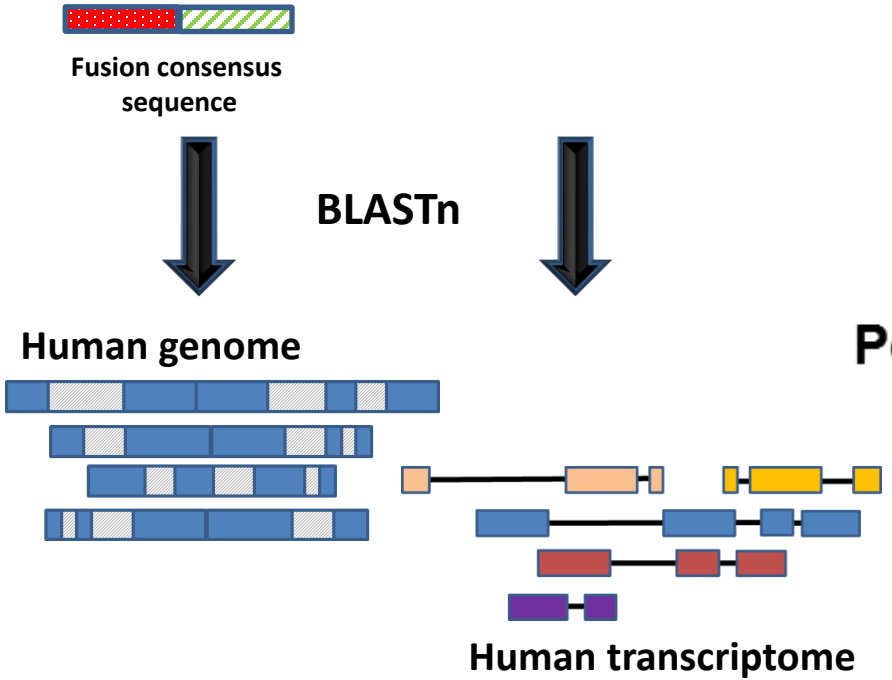
Filtering / Prioritization

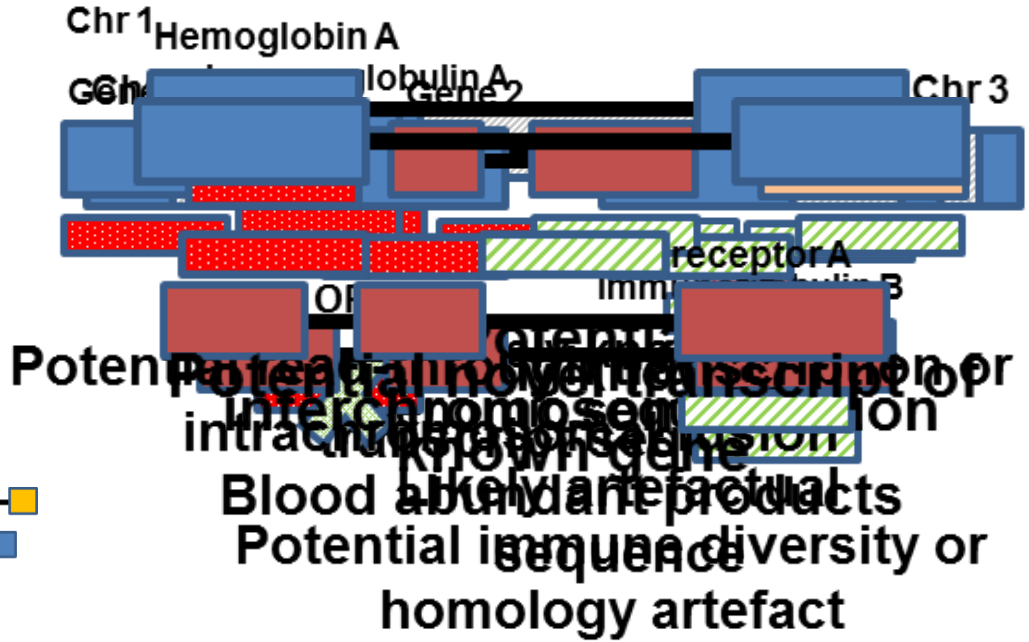
BLAST categorization



- 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





Now let's try it…

Questions