Structural Variation Classes

- **DELETION**
  - Autism, mental retardation, Crohn’s

- **NOVEL SEQUENCE INSERTION**
  - Alu/L1/SVA
  - Haemophilia

- **MOBILE ELEMENT INSERTION**
  - Alu/L1/SVA
  - Haemophilia

- **TANDEM DUPPLICATION**
  - Schizophrenia, psoriasis

- **INTERSPERSED DUPLICATION**

- **INVERSION**

- **TRANSLOCATION**
  - Balanced rearrangements
  - Chronic myelogenous leukemia

CNV: Copy number variants
Structural variation discovery with NGS data

- SVs: genomic alterations > 50 bp.
  - Databases:
    - DGV: http://projects.tcag.ca/variation/
  - Input: sequence data and reference genome
  - Output: set of SVs and their genotypes (homozygous/heterozygous)
  - Often there are errors, filtering required
  - SV detection methods can be based on statistical analysis or combinatorial optimization
  - Tools: VariationHunter, BreakDancer, MoDIL, CommonLAW, Genome STRiP, Spanner, HYDRA, etc.
Challenges

- Most SVs are embedded within or around segmental duplications or long repeats
  - If you use unique mapping, you will lose sensitivity
  - Ambiguous mapping of reads will increase false positives
  - Reference genome is incomplete; missing portions are duplications which cause more problems in accurate detection
- Many SVs are complex; many rearrangements at the same site
- CNV discovery is heavily studied but still not perfect; detection of balanced rearrangements are still problematic
Duplications and CNV hotspots

(A) Gaucher disease
(B) Familial juvenile nephronophthisis
(C) Fascioscapulohumeral muscular dystrophy
(D) Spinal muscular atrophy
(E) Congenital adrenal hyperplasia III
(F) Williams-Beuren syndrome
(G) Glucocorticoid-remediable aldosteronism
(H) Prader-Willi syndrome
(I) Angelman syndrome
(J) Polycystic kidney disease
(K) Charcot-Marie tooth disease (CMT1A)
(L) Hereditary neuropathy with liability to pressure palsies
(M) Smith-Magenis syndrome
(N) Neurofibromatosis
(O) Pituitary dwarfism

Human genome

Bailey et al., Science, 2002
Duplications: inter & intra

- 51,599 pairs of SDs
  - 18,559 pairs intrachromosomal
  - 32,740 pairs interchromosomal
- Non-redundant corresponds to 166 Mb (~5% of genome)

Bailey et al., Science, 2002
# Genome-wide SV Discovery Approaches

## Hybridization-based
- Iafrate et al., 2004, Sebat et al., 2004

## Sequencing-based
- Read-depth: Bailey *et al.*, 2002
- Sanger sequencing: Mills *et al.*, 2006

## Single molecule analysis
- Optical mapping: Teague *et al.*, 2010

## Hybridization-based Sequencing-based

## Single molecule analysis
- Optical mapping: Teague *et al.*, 2010

## 1000 Genomes Project
Detection diversity

Gains & Losses > 5 Kbp in the same 5 individuals

Fosmid clone
End-sequence pair
Kidd et al., 2008
(N = 1,206)

Ultra-dense tiling
array CGH
Conrad et al., 2010
(N = 1,128)

Affymetrix 6.0 SNP microarray
McCarroll et al., 2008 (N = 236)

Kidd et al. Cell, 2010
Sequence signatures of structural variation

- **Read pair analysis**
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size
- **Read depth analysis**
  - Deletions and duplications only
  - Relatively poor breakpoint resolution
- **Split read analysis**
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution
- **Local and de novo assembly**
  - SV in unique segments
  - 1bp breakpoint resolution
All these first algorithms used Sanger sequence, but laid out the basic principles for NGS analysis.
Read depth based algorithms

- Assume random (Poisson) distribution in read depth
- Multiple mapping:
  - WSSD (whole genome shotgun sequence detection)
- Unique mapping:
  - High(er) resolution: CNVnator, EWT (RDXplorer)
Read depth analysis: WSSD

- Uses database of random reads to confirm duplicated nature of the sequence
  - increased # of copies => increased number of reads
  - decreased # of copies => decreased number of reads
- Compute depth-of-coverage in 5kb windows (sliding by 1kb); select regions with increased depth as **duplications**, regions with reduced depth as **deletions** (WSSD method)

---

**Sequence to Test**

**Random Genome Sample**
(Whole-Genome Shotgun Sequence)

---

**deletion**  **unique**  **duplicated**

---

Bailey et al., Science, 2002
Multiple vs. unique mapping

Modified from Chiang & McCarroll, Nat Biotech, 2009
Read depth - Copy number correlation

$R = 0.92$

Alkan et al., Nature Genetics, 2009
WSSD: next-gen

- NGS specific problems
  - Short reads: MegaBLAST is replaced by mrFAST / mrsFAST
  - Common repeats: all repeats need to be masked
  - GC % bias needs to be fixed

- Improvement
  - Absolute copy number detection in 1 kb non-overlapping windows
  - Genotyping highly identical paralogs

Alkan et al., Nat Genet, 2009
Read depth distribution

- Read depth doesn’t really follow Poisson distribution
  - Biases against high and low GC %
GC\% correction: LOESS

\[ y' = y - c(x) \]

\[ c(x) = f(x) - e(x) \]
GC% correction (modified LOESS)

\[ k_{gc} = \frac{\mu_{total}}{\mu_{gc}} \]

\[ d'_{gc} = d_{gc} k_{gc} \]

The version in SegSeq and CNVnator
GC% correction

Before GC correction

After GC correction
Repeatmask reference

Map reads mrFAST/mrsFAST

Remove outliers & apply LOESS

Calculate read depth 1 kb windows

Remove outliers until the RD distribution is Poisson

Calculate copy number: \( CN = \frac{RD}{RD_{\text{avg}}} \)

Alkan et al., Nat Genet, 2009
Differentiating Paralogous Genes

Associated with psoriasis and Crohn’s disease

CFHR

Associated with color blindness

opsin

Alkan et al., Nature Genetics, 2009
Singly Unique Identifiers (SUNs)

Copy 1  ATACGGCATATAATATCCGACGATACATGATGTTAG
Copy 2  ATGCTAGGCGATGTAATATCCGACGACATACATGATGTTAG
Copy 3  ATACGGCATATAACATCCGACGATACATGATGTTAG
Copy 4  ATGCTACGATATAATATCCGACGATACATGATGTTAG
Copy 5  ATGCTACGATATAATATCCGACGATACATGATGTTAG
Copy 6  ATACGGCATGTAATATCCGACGATACATGATGTTAG

Sudmant et al., Science, 2010
Event-Wise Testing (EWT)

- Unique mappings are used
- No masking
- Window size 100 bp
- Probabilistic analysis

Yoon et al. Genome Research, 2009
Event-Wise Testing (EWT)

- Read counts are converted to Z score:
  - \( z_i = \frac{(RC_i - \mu_i)}{\sigma_i} \)
- Upper and lower tail probabilities
  - \( p_i^U = P(Z > z_i) \)
  - \( p_i^L = P(Z < z_i) \)
- Unusual events for interval \( A \), \( l = |A| \); \( L \) number of windows in chromosome; FPR: false positive rate

\[
\max\left\{ p_i^U \middle| i \in A \right\} < \left( \frac{FPR}{L/l} \right)^1 \\
\max\left\{ p_i^L \middle| i \in A \right\} < \left( \frac{FPR}{L/l} \right)^1
\]

Yoon et al. Genome Research, 2009
Unique mappings
Mappings with low MAPQ are discarded
Partitioning is based on mean-shift technique developed for image processing

Abyzov et al. Genome Research, 2011
CNVs with exome sequencing

- Exome sequencing: capture only coding exons from DNA and sequence
  - 1% of total genome
  - Good for protein coding variants but misses regulatory sequence, introns, etc.

- Whole genome sequencing generates random data, but exome does not

- Capture efficiency changes for every exon (n~200,000)

- CNVs from exons: ExomeCNV
Open problems (read depth)

- Deletions are the most studied, but still not perfect:
  - Many FPs and FNs
  - Breakpoint resolution is often poor
  - Different algorithms capture different CNVs
  - Overlap with other experimental methods is poor
- Duplications are studied in lesser detail
- Exome read depth analysis
  - Very poor results due to differences in capture efficiency
NEXT: READ PAIRS + SPLIT READS