Genome Sequencing and Structural Variation (2)

Analysis of matepairs for the identification of variants

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #11

Today

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants
Empirical
CDF
KS
MoDIL

- In the previous lecture we discussed structural variation (SV) and basic strategies for identifying SV by whole-genome sequencing
- We discussed read-depth analysis in some detail
- Today, we will concentrate on an algorithm that exploits information in readpairs for variant calling of insertions and deletions
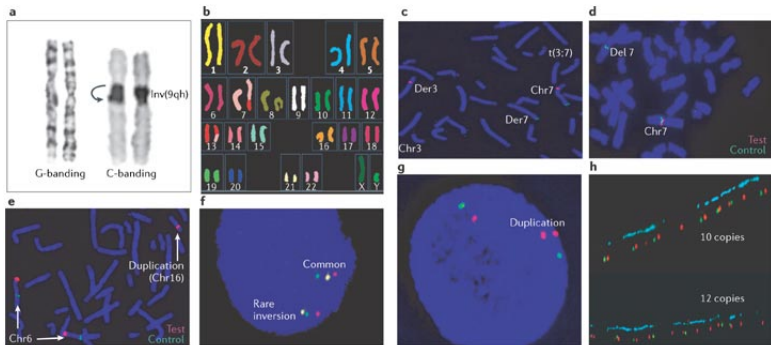
Structural variants

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Feuk L, et al. (2006) Structural variation in the human genome. *Nat Rev Genet* **7**:85-97.

Paired-end sequencing
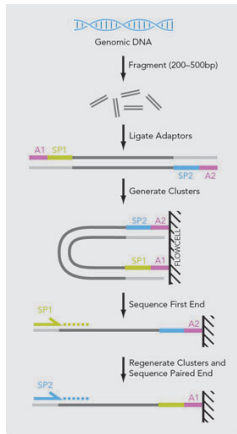
- Paired-end sequencing
- Sequence both ends of a single fragment with insert length typically 200–500 bp
- Typical paired end run on an Illumina GAIIx can achieve 2×75 bp reads and up to 200 million reads.

Graphic credit: Illumina

Matepair sequencing

WGS & SVs
(2)

Peter N.
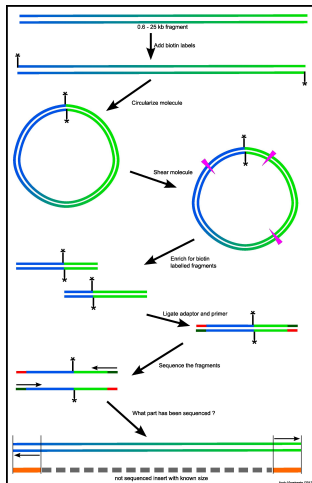Robinson

**Structural
Variants**

Empirical
CDF

KS

MoDIL

- Matepair sequencing
- Genomic libraries from the terminal sequences of genomic fragments that are of a uniform length (e.g. sequence 25-50bp terminal sequences of 3kb genomic fragments).
- When both tags of the "mate-pair" are independently mapped, they should be end up being the expected uniform distance apart (ie 3kb).

Graphic credit: Andry Vierstraete, Ghent
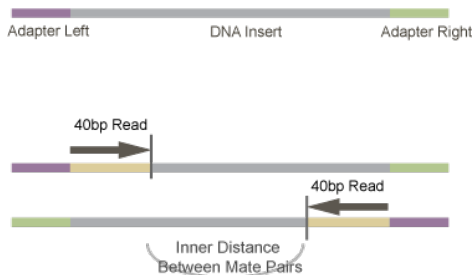
Matepair Nomenclature: Insert size

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- The nomenclature for paired-end or matepair reads

`Adapter(L)|-----40----|----------------------170----------------|-----40-----|Adapter(R)`

If we have two 40 bp paired-end reads with a 170bp middle piece, the insert size is calculated as

$2 \times 40 + 170 = 250$ nt. The fragment size is insert size plus length of both adapters ($\approx$ 120 nt extra).
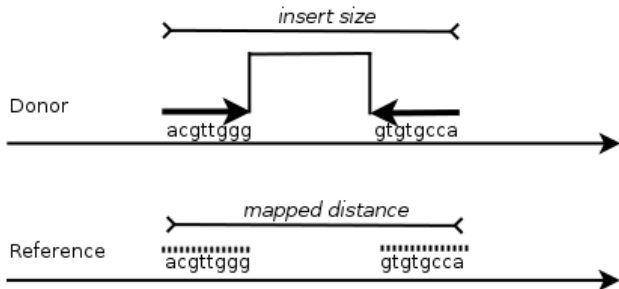
Matepair sequencing

- matepair with no structural variant
- The insert size is identical with the mapped distance between the paired sequence reads
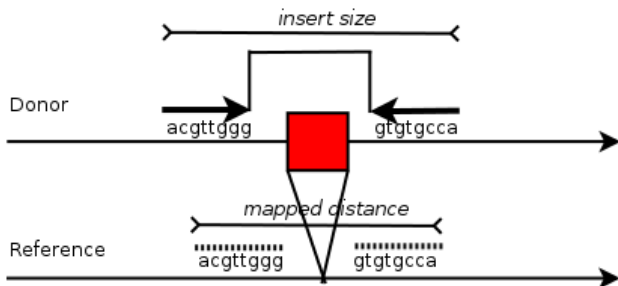
Matepair sequencing

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- matepair containing an insertion
- Since the reference doesn't have this insertion, the paired reads mapped closer to one another on the reference sequence than one would expect based on the insert size
- **Insertion size = insert size - mapping distance**
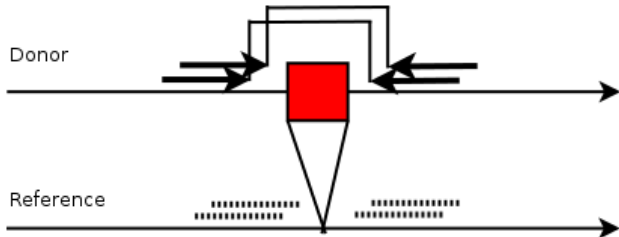
Matepair sequencing

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- Consider now two *overlapping* reads, X_i and X_j that contain the same insertion
- There is a consistent effect on mapping distance
- insert-size$_i$ − mapping-distance$_i$ ≈ insert-size$_j$ − mapping-distance$_j$

Matepair sequencing

- In reality, the mapping distance varies from readpair to readpair
- For a typical cluster of read pairs, the observed distribution of mapped distances is shown in grey
- This can be modeled by a Gaussian distribution with mean 208 bp and standard deviation 13 bp
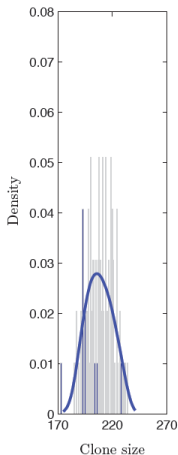
Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods* **6**:473–474

Matepair sequencing

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Today, we will examine a simplified version of MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions[a].

_____

[a] Lee S et al. *Nature Methods* **6**:473–474

To understand MoDIL, we will need to look at a few topics

- Empirical distribution function
- Kolmogorov Smirnov (KS) distribution
- Expectation-Maximization Algorithm (review)

Empirical distribution function

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Consider the following fundamental problem: Let $X_1, X_2, \ldots$ be an independent and identically distributed (iid.) sample from a distribution function F. Then, what does F "look like"?

We define the **Empirical distribution function**

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(X_i \leq x) \tag{1}$$

- **I** is the indicator function.
$$\mathbf{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in A \end{cases}$$

Cumulative distribution function

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

The empirical distribution function is a type of cumulative distribution function (CDF). A CDF describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x.

$$F_X(x) = P(X \leq x) \tag{2}$$

- The CDF thus represents the "area so far" function of the probability distribution.

Cumulative distribution function

- For a continuous random variable X, the CDF is defined in terms of its probability density function f as follows

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \qquad (3)$$

- For a discrete random variable X, the CDF is defined in terms of its probability density function f as follows

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i) \qquad (4)$$

CDF for Gaussian

WGS & SVs
(2)

Peter N.
Robinson

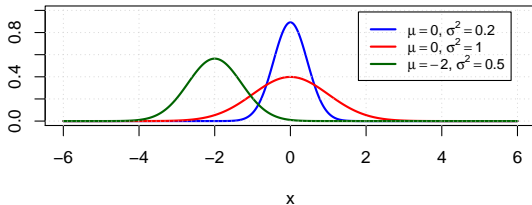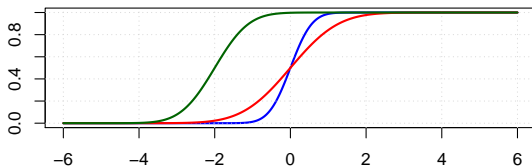Structural
Variants

Empirical
CDF

KS

MoDIL

Empirical CDF for Gaussian

Empirical Cumulative distribution function (5 points)

$\mu = 0, \sigma^2 = 1$

x
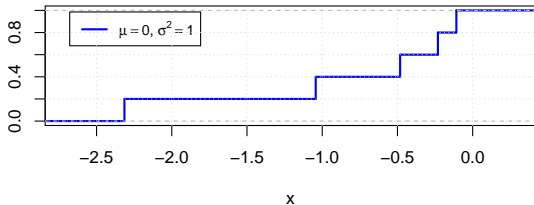
Empirical Cumulative distribution function (25 points)

$\mu = 0, \sigma^2 = 1$

CDF vs. empirical CDF

Thus, we have a dataset consisting of observed values of a random variable X, i.e., $x_1, x_2, \ldots, x_n$ consisting of observed values of random variables that are iid or independent identically distributed, that are iid or independent identically distributed

- There is a cumulative distribution function $F(x)$ representing the true, but potentially unknown distribution
- We estimate $F(x)$ using the empirical cumulative distribution function based on sampling from the distribution n times, $\hat{F}_n(x)$

Empirical CDF

- Consider now again the eCDF $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(X_i \leq x)$
- For a fixed value of x, the indicator function $\mathbf{I}(X_i \leq x)$ is a Bernoulli random variable with parameter $p = F(x)$
- Recalling that the binomial distribution corresponds to a sequence of Bernoullis, $n\hat{F}_n(x)$ follows a binomial distribution with parameters n and $F(x)$, i.e. $n\hat{F}_n(x) \sim \mathrm{binom}(n, F(x))$

Empirical CDF

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

**Empirical
CDF**

KS

MoDIL

Recalling that the expected value of a binomial random
variable X with $X \sim \mathrm{binom}(n, p)$ is $\mathbb{E}[X] = np$, we have that

$$
\begin{aligned}
\mathbb{E}[\hat{F}_n(x)] &= \frac{1}{n} \cdot \mathbb{E}[n \cdot \hat{F}_n(x)] \\
&= \frac{1}{n} \cdot \mathbb{E}[\mathrm{binom}(n, F(x))] \\
&= \frac{1}{n} \cdot nF(x) \\
&= F(x)
\end{aligned}
$$

Thus, $\mathbb{E}[\hat{F}_n(x)] = F(x)$, and thus the empirical CDF $\hat{F}_n(x)$ is
an unbiased estimator of the true CDF $F(x)$.

An alternative and more well known proof appeals to the law of large numbers

Kolmogorov–Smirnov statistic

The Kolmogorov–Smirnov (KS) statistic D_n compares a CDF with an empirical CDF

$$D_n = \sup_x |\hat{F}_n(x) - F(x)| \tag{5}$$

- Note that $\sup_x$ is the supremum (least upper bound) of x
- It can be shown that $\lim_{n\to\infty} D_n = 0$
- D_n can be used as the basis of a hypothesis test

Kolmogorov–Smirnov test

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Under the null hypothesis, $\sqrt{n}D_n$ converges to the Kolmogorov distribution

- The goodness-of-fit test or the Kolmogorov–Smirnov test rejects null hypothesis at level α if

$$\sqrt{n}D_n > K_\alpha \tag{6}$$

- The critical values of the KS distribution (K_α) are typically provided by the software (e.g., R) or can be looked up in tables.

We will not go into detail on the derivation of the KS theorem. If interested, a good place to start looking is

Doob JL (1949) Heuristic Approach to the Kolmogorov-Smirnov Theorems *Ann. Math. Statist* **20**:393–403,

which you can find online, as well as the Wikipedia entry

Kolmogorov–Smirnov statistic

Let us now explore how we might use the KS test for investigating insertions or deletions in WGS with paired-end or mate pair sequencing.

- Imagine we have the following cluster of read pairs that have been mapped to the reference genome
- The mapping distance for some is shorter than average (red), and for others is longer than average (blue)

Kolmogorov–Smirnov statistic
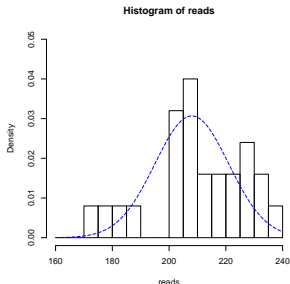
Let's imagine we have a distribution of insert distances for a cluster of readpairs. Based on the genome-wide distribution of insert distances, we have estimated the mean for the insert distance of a cluster of reads as 208 bp with a standard deviation of 13bp.

Histogram of reads

Kolmogorov–Smirnov statistic

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

We will start off by comparing the ECDF from 25 reads drawn at random from the true distribution with the actual distribution

```
> mu <- 208
> s  <- 13
> reads <- rnorm(25,mean=mu,sd=s)
> ks.test(reads, "pnorm", mean=mu, sd=s)
One-sample Kolmogorov-Smirnov test
data:  reads
D = 0.1371, p-value = 0.6848
alternative hypothesis: two-sided
```
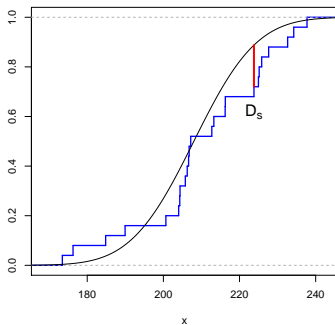
Kolmogorov–Smirnov statistic

- Thus, with a P-value of 0.6848 there is not enough evidence to reject the null hypothesis and we conclude that the insert distances are normally distributed according to $\mathcal{N}(\mu = 208, \sigma = 13)$.

Kolmogorov–Smirnov statistic

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- That is, we used the KS test to compare the observed distribution of insert lengths with that which we expect based on a Normal distribution, $\mathcal{N}(\mu = 208, \sigma = 13)$.

- The maximum distance between the two distributions of $D = 0.1371$ is the KS statistic, and we used the R command ks.test to perform the Kolmogorov–Smirnov test.

- The non-significant *p*-value of 0.6848 indicates that this test gave us no evidence to reject the null hypothesis that the two distributions are identical

Kolmogorov–Smirnov statistic

We will now imagine we have a homozygous deletion of 24bp. Then, on average readpairs will be mapped with at a distance of 24bp more than if there were no deletion. This will have the effect of shifting the mean to $\mu = 208 + 24 = 232$, without changing the standard deviation.
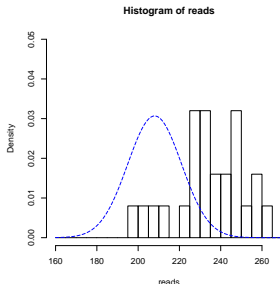
Histogram of reads

Kolmogorov–Smirnov statistic

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Performing the same ks test as above now yields

```
> ks.test(reads, "pnorm", mean=mu, sd=s)

One-sample Kolmogorov-Smirnov test

data:  reads
D = 0.74, p-value = 1.421e-14
alternative hypothesis: two-sided
```

Kolmogorov–Smirnov statistic
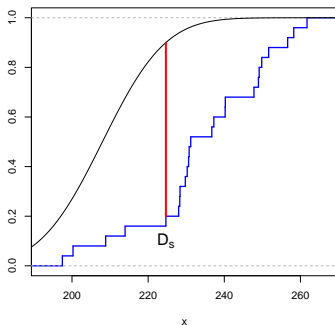
WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- The two curves, as well as D_s are much more clearly separated than in the first case

But why the KS test?

You may now be asking yourself why we should go to so much trouble to compare data which is normally distributed with a Normal distribution? Why not use a z-score or a t-test?

- You would, of course, be perfectly correct
- We have shown the above slides only to demonstrate how the KS test works in general.
- The KS test is a non-parametric test, meaning that it does not make many assumptions about the distribution of the data
- In contrast to the z-score or the t-test, the KS test can be used with data that are not (even close to being) normally distributed

Matepair sequencing

WGS & SVs
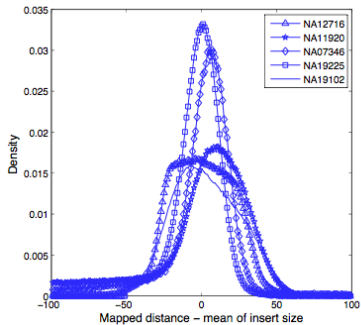(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

In fact, the empirical distribution of readpair mapped distances is not Gaussian and it differs from sequencing library to sequencing library

Two Sample KS Test

Therefore, because there is no analytic distribution that we can use to model the distribution of insert lengths, we will use the two-sample Kolmogorov–Smirnov test, which is used to test whether two one-dimensional probability distributions differ.

- We let $F_1(x)$ denote the first empirical CDF (with data from across the entire genome), and $F_2(x)$ be the second one (with data from the cluster of interest)
- The two sample KS test is then defined as

$$D_{n_1,n_2} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)| \qquad (7)$$

- F_{1,n_1} and F_{2,n_2} are the empirical distribution functions for samples 1 and 2 (with sample sizes n_1 and n_2)

Two Sample KS Test

The null hypothesis is rejected at significance level α if

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1, n_2} > K_\alpha$$

- Tables are available in statistical software for the values of K_α
- For the two-sample KS test, K_α can be approximated by

$$K_\alpha = c(\alpha) \frac{n_1 + n_2}{n_1 n_2}$$

e.g., for $\alpha = 0.05$, $c(\alpha) = 1.36$, for $\alpha = 0.01$, $c(\alpha) = 1.63$

Two Sample KS Test

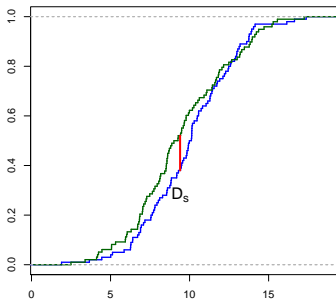- A: $\mathcal{N}(\mu = 10, \sigma = 3)$
- B: $\mathcal{N}(\mu = 9.5, \sigma = 3.2)$
- $D_s = 0.1733$
- p-value = 0.08707
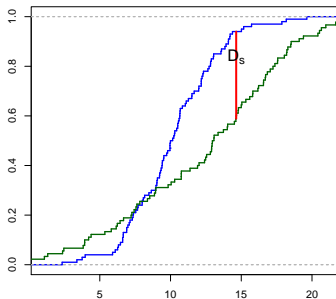
- A: $\mathcal{N}(\mu = 10, \sigma = 3)$
- B: $\mathcal{N}(\mu = 12, \sigma = 5)$
- $D_s = 0.3133$
- p-value = 0.0001273

MoDIL

With all of this in hand, we can now examine the MoDIL algorithm:

Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: detecting small indels from clone-end sequencing

with mixtures of distributions. *Nature Methods* **6**:473–474

Bird's eye view

1. Cluster matepairs
2. Estimate distribution of insert sizes across genome: $p(Y)$
3. For each individual cluster C_i, check by EM algorithm whether the distribution of insert sizes corresponds to homozygous reference, heterozygous indel, or homozygous indel
   - Expectation: What haplotype does each read of C_i belong to?
   - Maximization: Optimize KS statistic to estimate μ_1 and μ_2

1) Cluster matepairs

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL
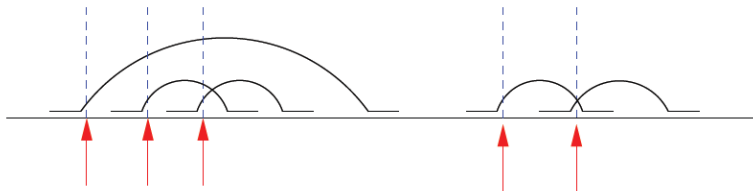
- Red arrows: genomic locus one bp after the left read
- For each red arrow, all matepairs spanning the genomic location (blue dotted line) are defined as the cluster corresponding to that genomic location

- Arrows # 1 & # 4: one matepair; arrows #2 and # 5: two mate pairs; arrow # 3: three matepairs

1) Cluster matepairs
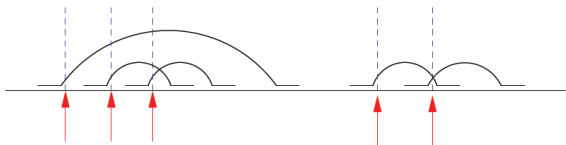
Algorithm 1 Cluster matepairs

1: **for** each mapped location X_i **do**

2:    Find all $\{X_j\}_{j=1}^{L}$ with overlap to X_i

3:    Define cluster $C_k = \{X_1, X_2, \ldots, X_L\}$

4: **end for**

- The mapped locations X_i correspond to the arrows in the Figure

- Note this algorithm allows closely located clusters to share matepairs

2) Estimate genomewide distribution of insert size

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- Use all readpairs with consistent mappings to estimate the mean and the standard deviation of the insert size
- Call this distribution $\boxed{p(Y)}$. Note that $p(Y)$ is not Gaussian and we will use the empirical CDF
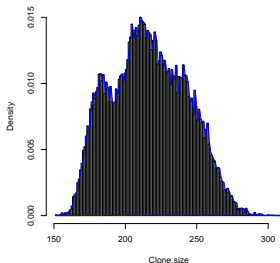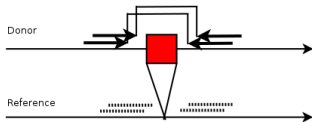
genome wide peak distribution (deviously simulated)

2) Estimate genomewide distribution of insert size: $p(Y)$

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS
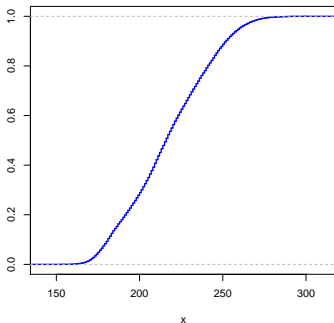
MoDIL

- The empirical CDF of insert lengths is calculated as above
- Shown here for the simulated reads from the previous slide

Comparing $p(C_i)$ to $p(Y)$ with no indel
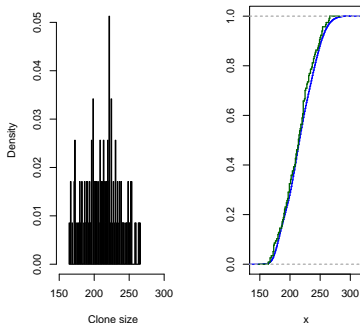
- On the left we see a density plot for 117 reads of a simulated cluster drawn at random from the genomic insert length distribution.

- The 2-sample KS test yields: $D = 0.073$, p-value $= 0.5637$

Comparing $p(C_i)$ to $p(Y)$ with 24 bp deletion

WGS & SVs
(2)

Peter N.
Robinson
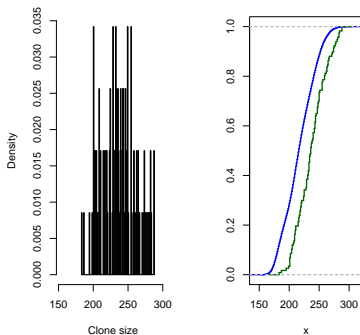
Structural
Variants

Empirical
CDF

KS

MoDIL

- On the left we see a density plot for 117 reads of a simulated cluster drawn at random from the genomic insert length distribution shifted by 24bp (homozygous deletion).
- The 2-sample KS test yields: 0.3121, p-value $= 2.77 \times 10^{-10}$

Estimating size of indel events
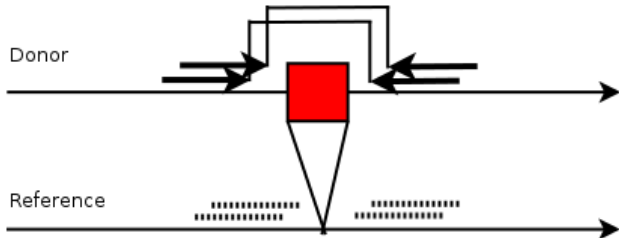
The expected size of an indel follows a Gaussian distribution with mean $\mu_{p(y)} - \mu_{p(C_i)}$ and standard deviation $\sigma = \dfrac{\sigma_{p(C_i)}}{\sqrt{n}}$

Estimating size of indel events

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Lemma

If y is the mean of the insertsize in the entire library and X_i is the mapped insert size for readpair i in some cluster C_j, then the random variable $Z_i = y - X_i$ represents the indel size corresponding to readpair i. Then $Z_1, Z_2, \ldots, Z_n$ are random variables with mean μ_Z and standard deviation σ_Z. If there are n mate pairs in the cluster, then the expected value of their mean is Gaussian distributed and the distribution of the sample average has a standard deviation of $\frac{\sigma_Z}{\sqrt{n}}$

Estimating size of indel events

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Proof.

The fact that the sample mean converges to μ_Z is a consequence of the law of large numbers, and the central limit theorem implies that the variance of the sample mean will tend to $\frac{\sigma_Z^2}{n}$ $\qquad\square$

- Thus as n, the number of readpairs in the cluster, grows, our confidence in the size of the indel increases
- Smaller indels can be predicted the higher the coverage is

But...the human genome is diploid

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants
Empirical
CDF
KS
MoDIL

- During sequencing it is typically impossible to distinguish between the matepairs coming from each of the chromosomes.
- Thus clusters actually consist of matepairs from both haplotypes.
- If the observed cluster is the site of a homozygous indel both of the distributions will shift simultaneously.
- If, however, the indel is heterozygous approximately half of the observed matepairs will be generated from the shifted distribution. while the other half will come from the original, unshifted $p(Y)$.

But...the human genome is diploid

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS
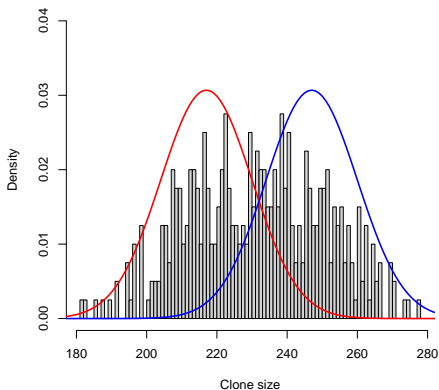
MoDIL

- red: no indel; blue: deletion of 30 bp

Mixture of distributions

- To model the fact that there are two chromosomes (haplotypes) from which the mapped paired reads can come, MoDIL used two random variables to model the expected indel size (one for each haplotype)

- This is a mixture of distributions (MoD): Not of Gaussian or binomial distributions as we have seen previously, but of empirical CDFs (the empirical distribution of clone sizes, $p(Y)$)

- This is a classic application for expectation maximization-type algorithms

Mixture of distributions

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- Given a cluster C_i, MoDIL identifies the two distributions which have the fixed shape of $p(Y)$ and arbitrary means that best fit the observed data using the Kolmogorov-Smirnov (K-S) goodness of fit test
- If there is no indel, we expect $\mu_\alpha = \mu_\beta = \mu_{p(Y)}$
- if there is a homozygous indel, we expect
  $\mu_\alpha = \mu_\beta = \mu_{p(C_i)} \neq \mu_{p(Y)}$
- If there is a heterozygous indel, then $\mu_\alpha = \mu_{p(Y)}$ and
  $\mu_\beta = \mu_{p(C_i)} \neq \mu_{p(Y)}$
- The means of the two distributions are found using the Expectation- Maximization algorithm.

EM: The hidden variables

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

We cannot observe which haplotype a given readpair comes from

- Previously, we defined $Z_i = y - X_i$ for the expected indel size
- In the diploid situation, we instead let Z_i be represented by two hidden variables, Z_i^α and Z_i^β, with m of the readpairs being from Z_i^α and n of the readpairs from Z_i^β
- If we somehow could observe which haplotype a given readpair came from, it would be trivial to estimate the indel size according to Lemma 1

EM: The hidden variables

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- Now let $\mathbf{s} = \{\mu_1, \mu_2\}$ be the means of Z_i^{α} and Z_i^{β}
- Note that MoDIL assumes that Z_i^{α} and Z_i^{β} have the same shape as $p(Y)$, with the mean possibly being shifted, but the form of the empirical CDF otherwise being identical
- If we denote by γ_{jt} the likelihood that the jth readpair was generated from haplotype t, then we define π_t as

$$\pi_t = \frac{\sum_j \gamma_{jt}}{\sum_{t \in \{1,2\}} \sum_j \gamma_{jt}} \tag{8}$$

- π_t is thus the estimated proportion of reads coming from haplotype t[1]

---

[1] recall the mixture parameter from lectures #8/#9.

EM: The hidden variables

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- γ_{jt} is defined as the posterior probability that the jth readpair was generated from haplotype t

$$\gamma_{jt} = \frac{P(Z_j|\mu_t) \times \pi_t}{\sum_{t' \in \{1,2\}} P(Z_j|\mu_{t'}) \times \pi_{t'}}$$

- prior
- Likelihood
- Normalizing constant

EM: likelihood term

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

- But how do we calculate $P(Z_j|\pi_t)$? (Likelihood of insert length Z_j given haplotype t)?
- We have the empirical CDF that is shifted according to the value of μ_t
- It can be shown that the size of the jump at x for the empirical CDF is the value of the probability density function at x.
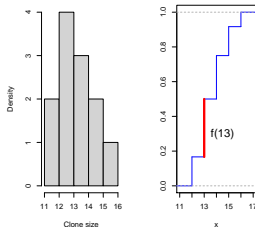
The probability of a read of length 13 is shown by the jump at $x = 13$ (red)

EM: Expectation Step

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

The E-step therefore estimates the responsibility of the two distributions for each readpair γ_{jt} using the current values for μ_1 and μ_2 using equation (9).

$$\gamma_{jt} = \frac{P(Z_j|\mu_t) \times \pi_t}{\sum_{t' \in \{1,2\}} P(Z_j|\mu_{t'}) \times \pi_{t'}} \tag{9}$$

EM: Maximization Step

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

The M-Step searches for an optimal $\mathbf{s} = \{\mu_1, \mu_2\}$ to minimize the sum of the Kolmogorov–Smirnov goodness of fit statistics for the two haplotypes as shown in equation (10).

$$D = \sum_{t \in \{1,2\}} \pi_t \sup_z |F_t^0(z) - F_t(z)| \qquad (10)$$

- $\sup_z |F_t^0(z) - F_t(z)|$ is the two-sample KS statistic, whereby F^0 is the empirical CDF of the genome-wide distribution and F_t is the empirical CDF for the cluster (whose mean may be shifted)
- π_t is the prior probability of a read belonging to haplotype t

EM: Maximization Step

$F_t^0(z)$ is the eCDF for z, weighted by the posterior that the read belongs to the haplotype t (γ_{jt})

$$F_t^0(z) = \frac{1}{\sum_i \gamma_{it}} \sum_{j=1}^{L} \gamma_{jt} \mathbf{I}(Z_j \leq z) \tag{11}$$

Recall that $Z_j = \mu_{p(Y)} - X_j$, where $\mu_{p(Y)}$ is the library-wide mean insert size

EM: Maximization Step

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Similarly, $F_t(z)$ is the eCDF for z according to the distribution of of $p(Y - y + \mu_t)$, i.e., the shifted distribution of $p(Y)$ with mean μ_t

$$F_t(z) = \frac{1}{\sum_i \gamma_{it}} \sum_{j=1}^{L} \gamma_{jt} \mathbf{I}(Z_j \leq z) \tag{12}$$

Recall that for the (potentially shifted) distribution of haplotype t of cluster i , $Z_j = \mu_t - X_j$, where μ_t is mean insert size for haplotype t of cluster i

EM: Maximization Step

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

In the M step, a range of possible indel sizes are evaluated in order to find the $\mathbf{s} = \{\mu_1, \mu_2\}$ that minimizes $D = \sum_{t \in \{1,2\}} \pi_t \sup_z |F_t^0(z) - F_t(z)|$.

- To search for insertions, the space $[-1000, y]$ is searched, where y is the mean library insert size.[2]
- 1000 is an arbitrary bound that may be altered depending on the experimental parameters
- To search for deletions, the space $[\mu_C - 100, \mu_c + 100]$ is searched, where μ_C is the mean of all the mapped distances in the cluster.

---

[2]Recall that an insertion reduces the mapping distance. Note that MoDIL cannot find insertions greater than the insert size.

EM: Maximization Step

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Algorithm 2 Expectation Maximization

1: Initialize $\mathbf{s} = \{\mu_1, \mu_2\}$ and $\pi_1 = \pi_2 = 0.5$.
2: **repeat**
3:   E step: estimate posterior probability γ_{jt}
     for each readpair using current parameters
     and Eq. (9)
4:   M step: update $\mathbf{s} = \{\mu_1, \mu_2\}$ by minimizing KS
     statistic according to Eq. (10); update π_1
     and π_2 by Eq. (8).
5: **until** KS statistic reaches (local) optimum
6: Discard cluster if KS test rejects null
   hypothesis[3]

---

[3] Then, the cluster insert sizes do not follow the expected distribution $p(Y)$ and likely represent an artifact

MoDIL: Bells and Whistles

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

While the above pages give the gist of the MoDIL algorithm, there are a number of heuristics and additional features that will be summarized at a bird's eye level here

- To initialize $\mathbf{s} = \{\mu_1, \mu_2\}$, consider that most clusters have no indel, so it is prudent to initialize to $\mathbf{s} = \{0, 0\}$

- The second most common case is a heterozygous indel. In this case, the size of the indel should be twice the sample mean of the expected size of indels, Z_j, which is

$$2\mu_C = \frac{2}{\sum_i \gamma_{it}} \sum_{j=1}^{L} \gamma_{jt} Z_j$$

- Thus in a second iteration, initialize to $\mathbf{s} = \{2\mu_C, 0\}$.

MoDIL: Bells and Whistles

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

A number of other heuristics are used to avoid overfitting problems, explore the search space, and deal with noisy data

- Indel events are weighted according to heterozygous > homozygous > mixed (two different indels)
- A factor graph is used to assign unique mapped locations to readpairs (e.g., a read pair may be mapped incorrectly due to repeats)
- Adjacent compatible clusters are merged
- Clusters are classified as heterozygous or homozygous

MoDIL: Evaluation

Table 1 | Evaluation of MoDIL on various datasets

Size	Type	MoDIL	Overlap with known indels[7]			Simulation	
			Total	Found	FNR	Recall	Precision
≥20 bp	Insertion	1,336	78	75	0.04	0.85	0.90
	Deletion	3,799	196	187	0.05	0.91	0.89
15–19 bp	Insertion	1,601	119	84	0.29	0.61	0.65
	Deletion	5,333	178	126	0.29	0.78	0.45
10–14 bp	Insertion	936	370	130	0.65	0.44	0.37
	Deletion	3,682	593	227	0.62	0.54	0.27

Number of insertions and deletions of each size identified by MoDIL from Illumina data[6] and the number of previously known indels[7] (total) overlapped by MoDIL predictions (found). We considered indels discovered in ref. 7 but not by us to be false negatives, and the ratio of these as a function of all indels in ref. 7 the false negative rate (FNR). Using a simulated dataset (simulation), we computed the fraction of true indels discovered by MoDIL (recall) and the fraction of predicted indels that were real (precision).

recall: fraction of true indels that were called

$$\text{Recall} = \frac{|\{\text{Called indels}\} \cap \{\text{True indels}\}|}{|\{\text{True indels}\}|} \qquad (13)$$

precision: fraction of called indels that truly are indels

$$\text{Precision} = \frac{|\{\text{Called indels}\} \cap \{\text{True indels}\}|}{|\{\text{Called indels}\}|} \qquad (14)$$

MoDIL: Evaluation

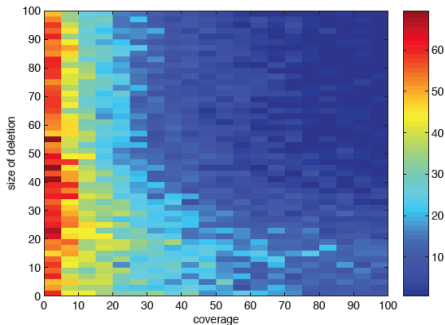WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Heatmap: power of MoDIL to detect heterozygous variants of various sizes at different coverage levels.

- dark red means 70% error/30% correct
- dark blue 0% error/100% correct.
- What does this tell us about our ability to identify indels of different sizes? Why do you think this is so?

Finally...

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Why do you think the authors of MoDIL use the KS goodness of fit test instead of maximum log likelihood for their version of the EM algorithm?
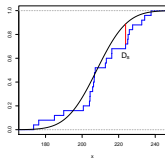
Finally...

Consider the effect of outliers . . .

Does it matter how extreme the lowest value, say, in the eCDF is? How does this effect our estimate of D_S?

Recall the definition of the **Empirical distribution function**

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(X_i \leq x)$$

$$\hat{\mu}_{MAP} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Maximum likelihood estimator for the mean of a Gaussian

Recall this expression for the maximum likelihood estimate of μ

from the lecture on SNV calling

What effect do individual extreme values (outliers) have here?

Finally...

The KS test was chosen instead of the maximum log likelihood because of its robustness to outliers.

- The log likelihood is sensitive to samples with low density in a distribution.

- For example, if Z_j is far away from the mean of the distribution, then it significantly reduces the log likelihood, and to minimize the effect, the mean of the distribution will be shifted toward this outlier. Thus, if there exist a few outliers in a cluster, it may falsely predict that there is an indel even though majority of samples suggest no indel.

- In contrast, the K-S statistic is largely influenced by samples with large density in the distribution, and is quite insensitive to outliers, which is a desirable property for noisy datasets.

Summary

WGS & SVs
(2)

Peter N.
Robinson

Structural
Variants

Empirical
CDF

KS

MoDIL

Important topics from this lecture

- cumulative distribution function
- empirical CDF
- Kolmogorov–Smirnov distribution/test
- Effect of indels on mapped insert size
- relation between number of read pairs in a cluster, size of an indel, and detectability

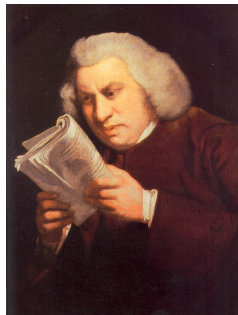# The End of the Lecture as We Know It

- Email: peter.robinson@charite.de
- Office hours by appointment

Lectures were once useful; but now, when all can read, and books are so numerous, lectures are unnecessary. If your attention fails, and you miss a part of a lecture, it is lost; you cannot go back as you do upon a book... People have nowadays got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the books from which the lectures are taken. I know nothing that can be best taught by lectures, except where experiments are to be shown. You may teach chymistry by lectures. You might teach making shoes by lectures!

Samuel Johnson, quoted in Boswell's Life of Johnson (1791).