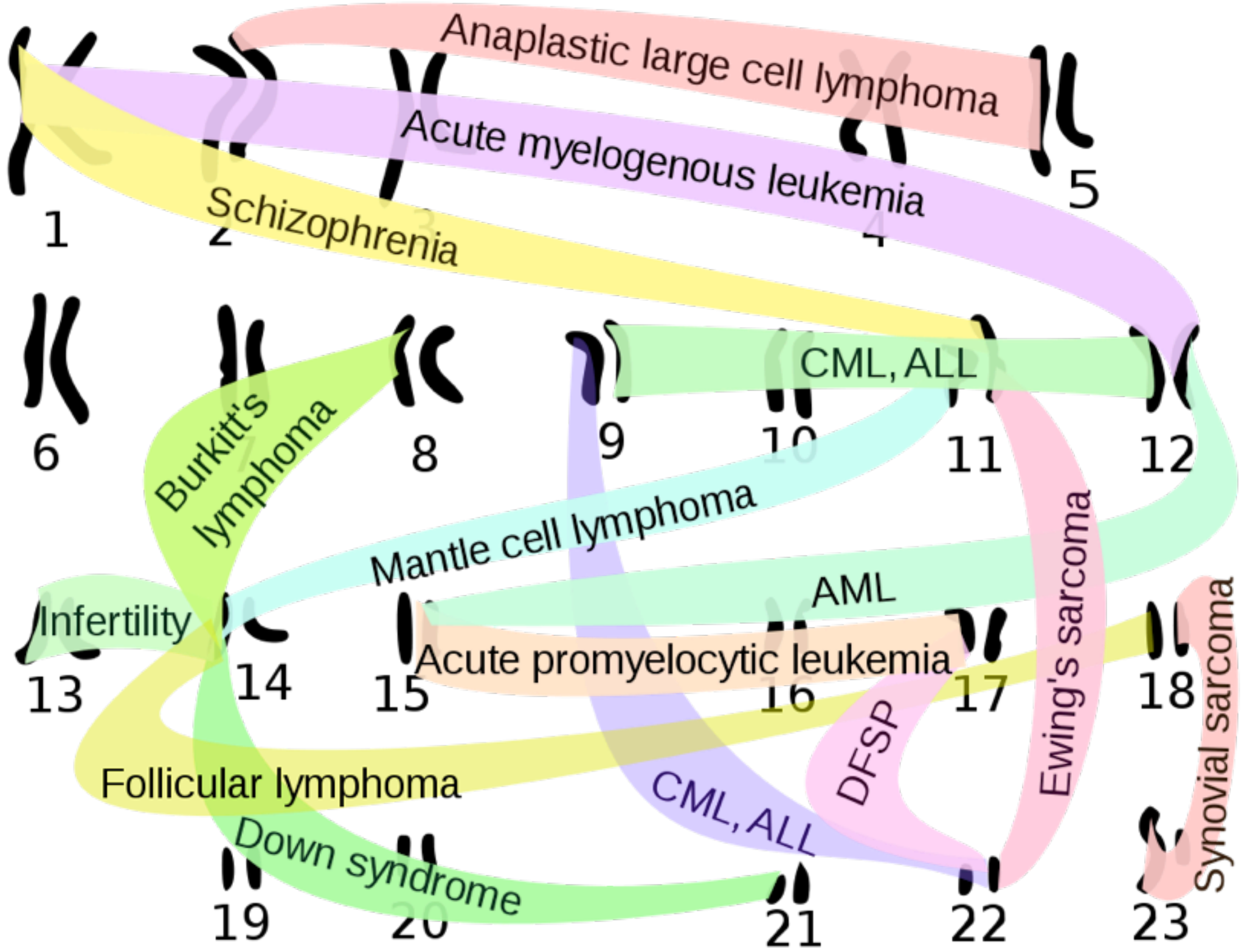
Metody detekcji i analizy patogennych zmian strukturalnych w genomie człowieka

Anna Gambin,
Instytut Informatyki,
Uniwersytet Warszawski

Wykład 1: testowanie hipotez statystycznych

- stabilność genomu, mutacje strukturalne.
- podstawowe testy statystyczne.
- testowanie hipotez dotyczących wpływu architektury genomu na jego stabilność.


Anaplastic large cell lymphoma
Acute myelogenous leukemia
Schizophrenia
1
2
3
4
5
6
7
8
9
10
11
12
CML, ALL
Burkitt's lymphoma
Mantle cell lymphoma
Infertility
AML
Acute promyelocytic leukemia
13
14
15
16
17
18
DFSP
Ewing's sarcoma
Synovial sarcoma
Follicular lymphoma
CML, ALL
Down syndrome
19
20
21
22
23

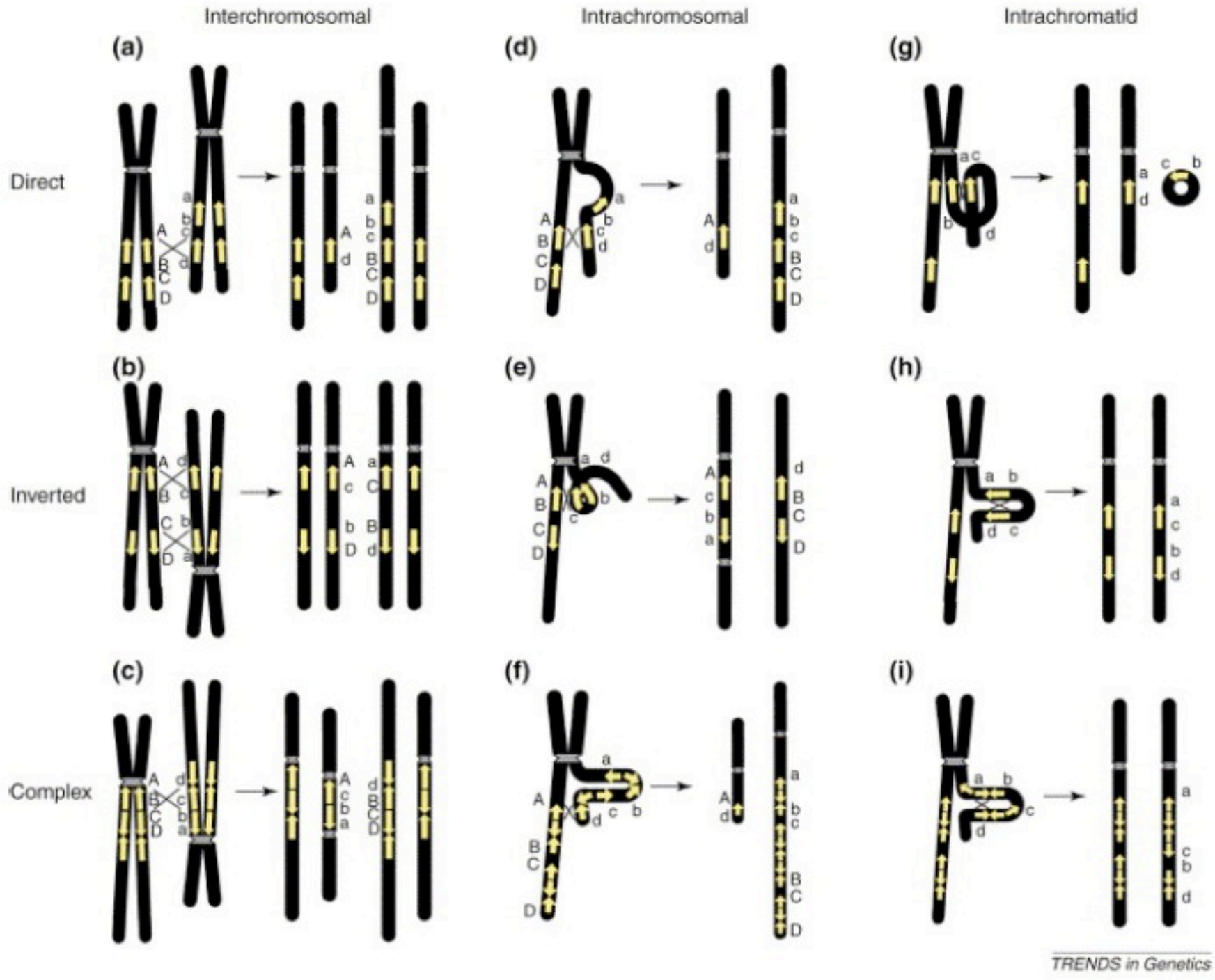
Zaburzenia genomowe

- wysokopoziomowe cechy architektury genomu moga zwiekszać podatność na rearanżacje DNA (zaburzenia genomowe, *ang. genomic disorders*) co może być przyczyna chób genetycznych
- interesuja nas rearanżacje *de novo* - czyli takie, które nie sa dziedziczone
- ważna przyczyna, która wpływa na zmiany w fenotypie jest różnica w liczbie kopii genów wrażliwych na dawke (*ang. dosage sensitive*) albo zaburzenia regionów regulatorowych tychże genów

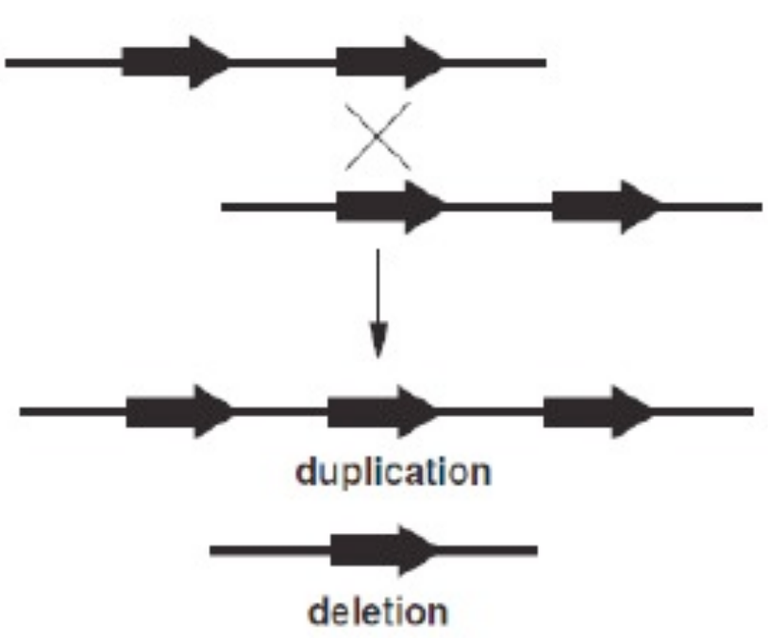
Powtórzone fragmenty sekwencji (LCRy)

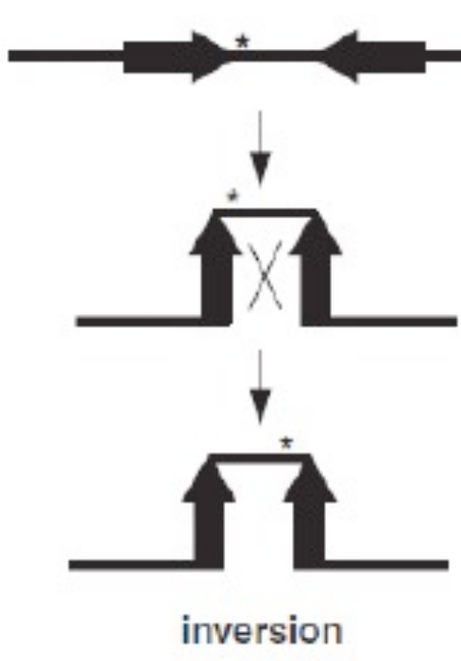
- ang. Low-Copy Repeats
- fragmenty DNA $>$ 1 kb oraz posiadajace $>$ 90% podobieństwa sekwencji (fraction matching)
- LCRy $>$ 10 kb oraz charakteryzujace sie podobieństwem sekwencji na poziomie $>$ ok. 97% moga powodować zaburzenia w genomie
- LCRy moga mediować mechanizm Non Allelic Homologous Recombination (NAHR)

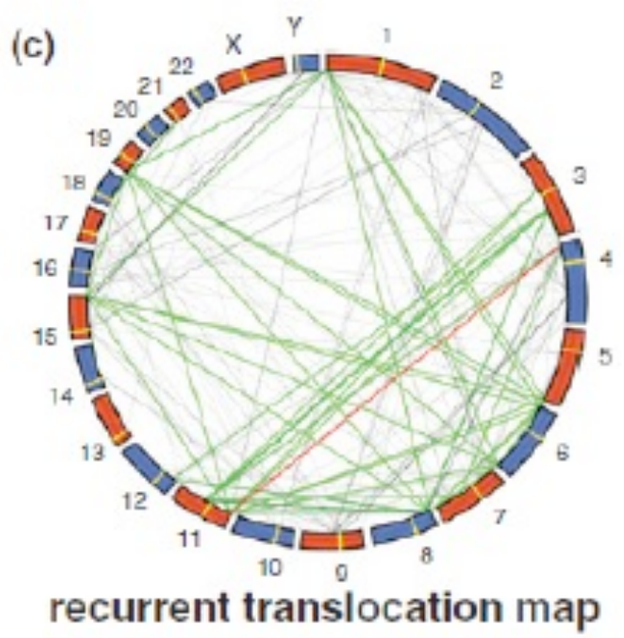
LCRs, NAHR oraz zaburzenia genomowe

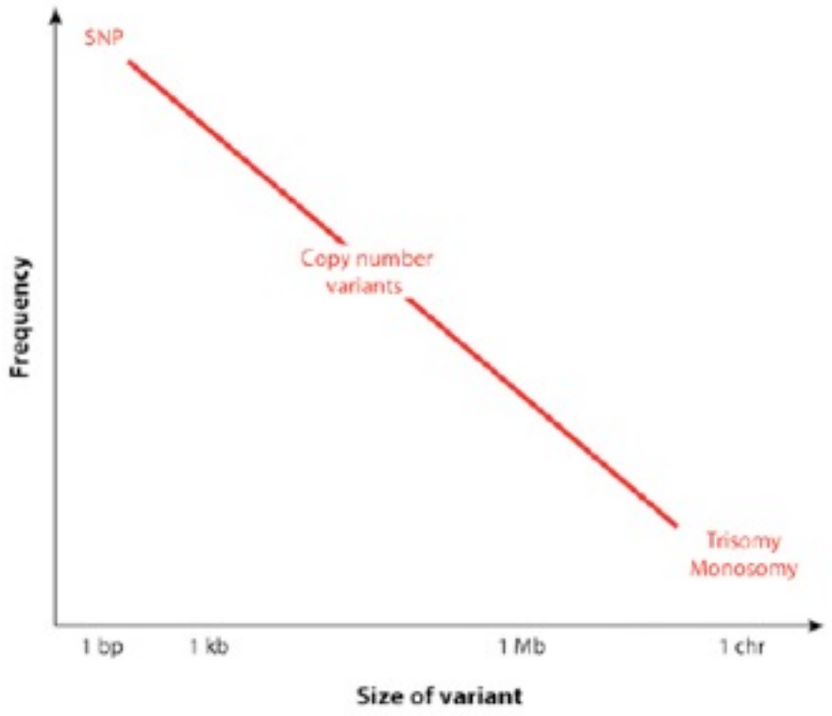


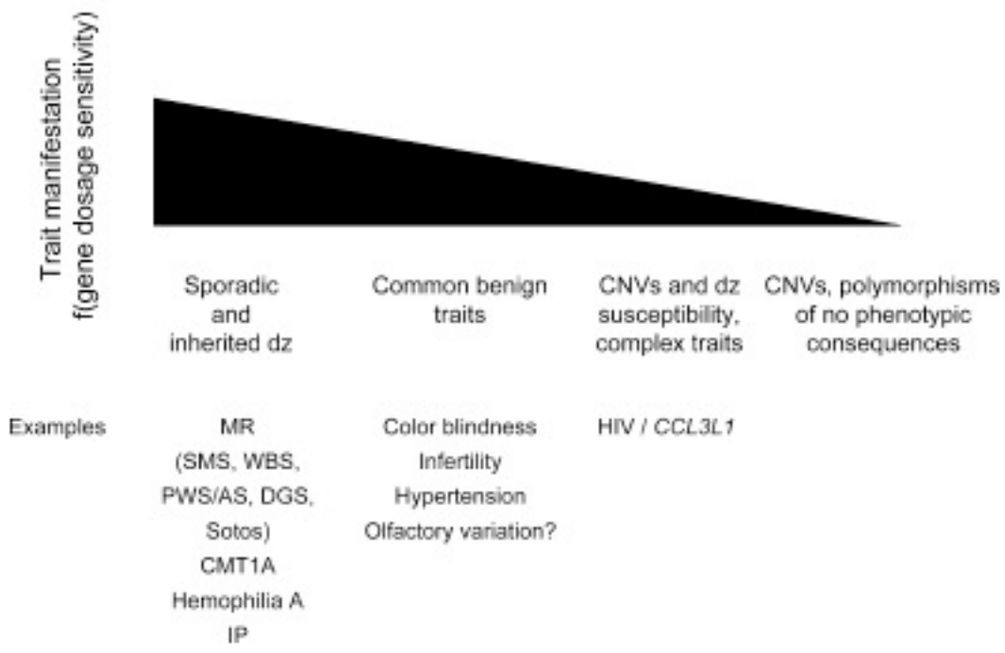
source: Stankiewicz et al., 2002


duplication
deletion


inversion

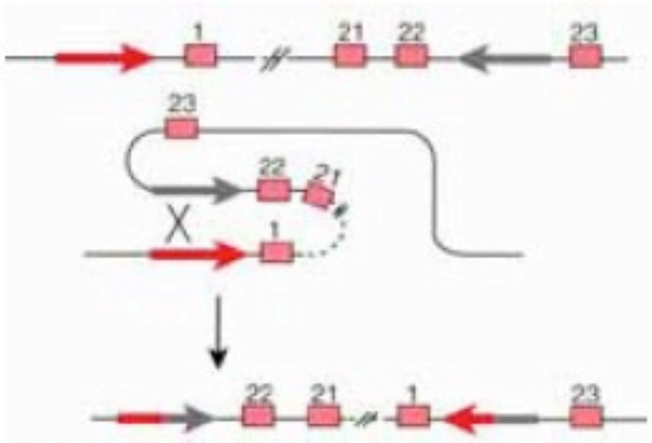

(c)
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
recurrent translocation map





Girirajan S, et al. 2011.
Annu. Rev. Genet. 45:203–26

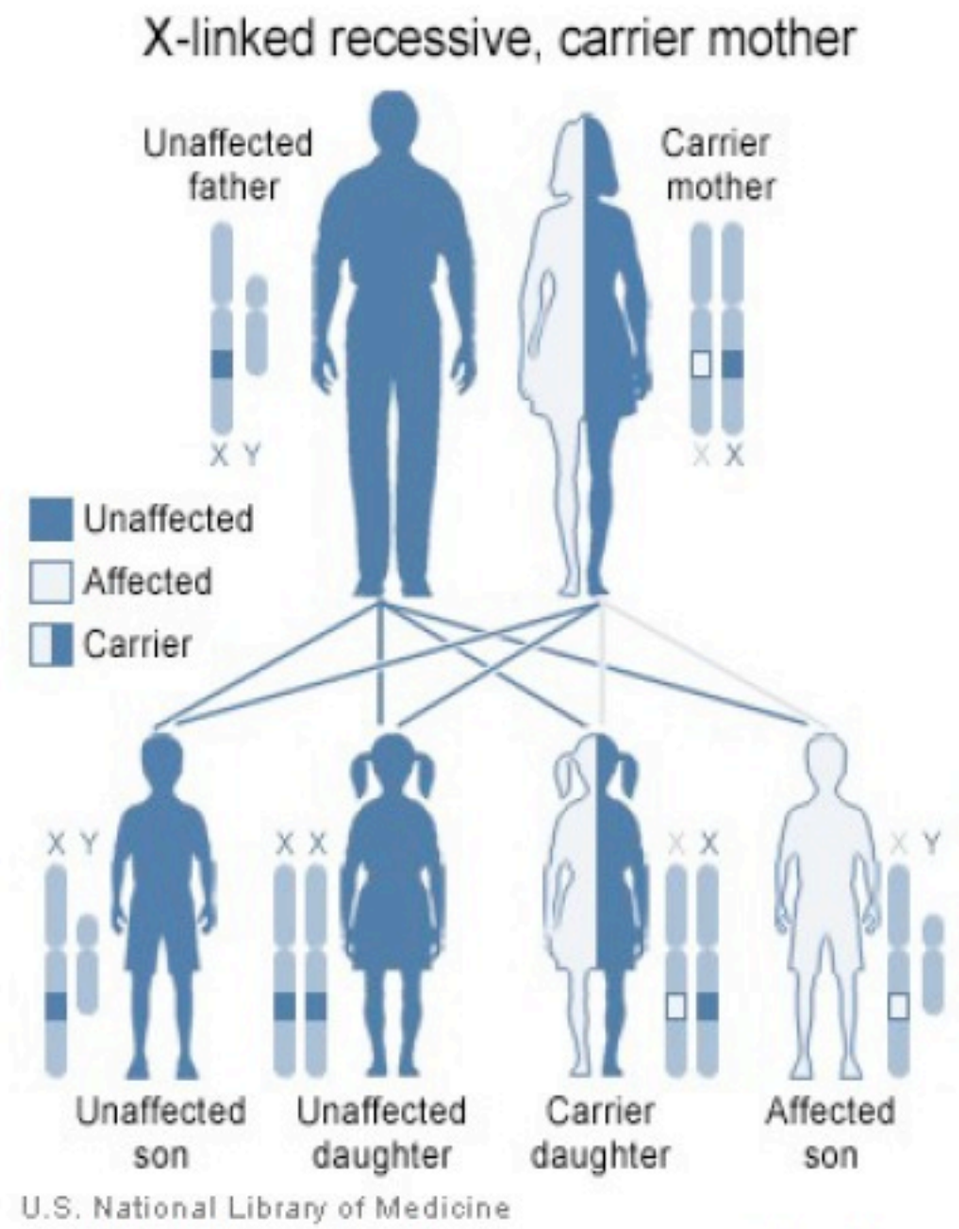
Heamophilia A



source: http://carolguze.com/text/442-2-mutations.shtml

- recessive X-linked genetic disorder
- impair the body's ability to control blood clotting or coagulation
- present in about 1 in 5,000-10,000 male births
- single inversion disrupting the factor VIII gene (F8) accounts for $>$ 45% of severe cases

X-linked disorders

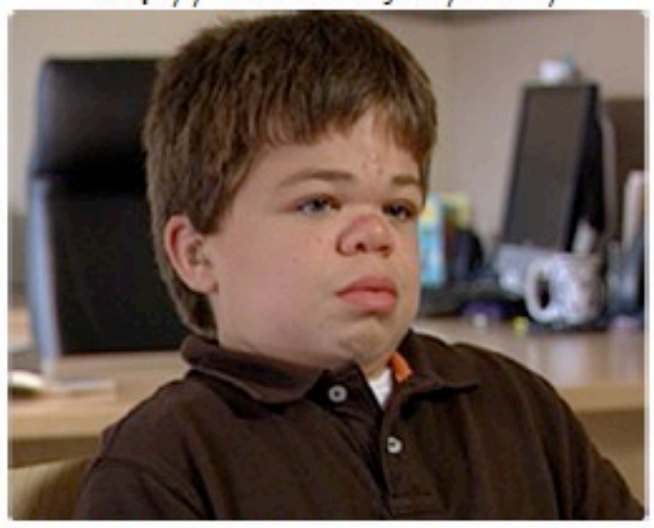


X-linked recessive inheritance, source: wikipedia

Hunter syndrome (mucopolysaccharidosis Type II)

- recessive X-linked genetic disorder
- impair the body's ability to control blood clotting or coagulation
- affects approximately 1 in 155,000 live male births (rarely reported to occure also in females)
- caused by a deficient (or absent) enzyme, iduronate-2-sulfatase (I2S)
- 13% of patients have the IDS gene disrupted by an NAHR-mediated inversion between the IDS gene and its pseudogene

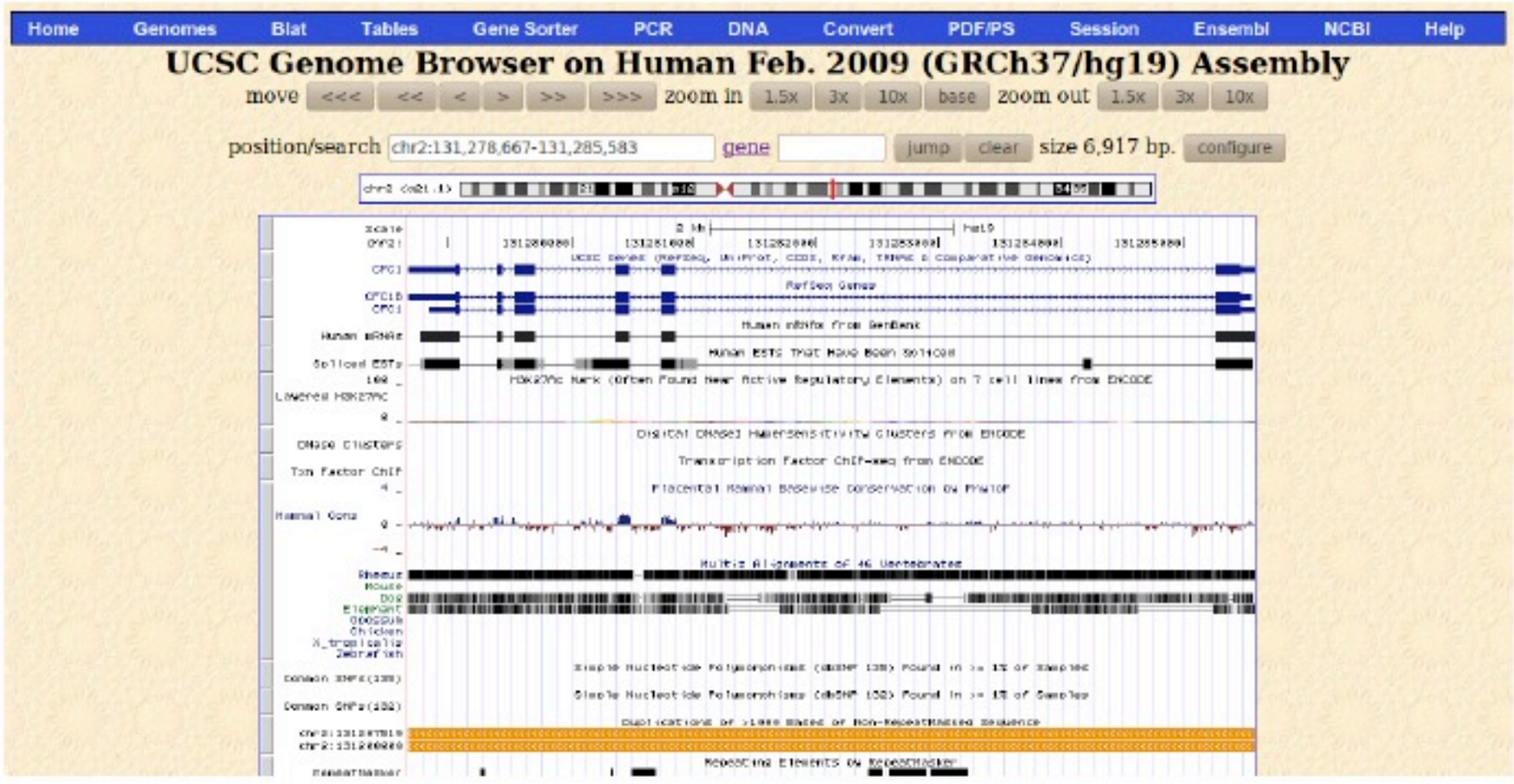
http://davechidley.ca/news/



http://www.hunterpatients.com/hunter-syndrome-community/

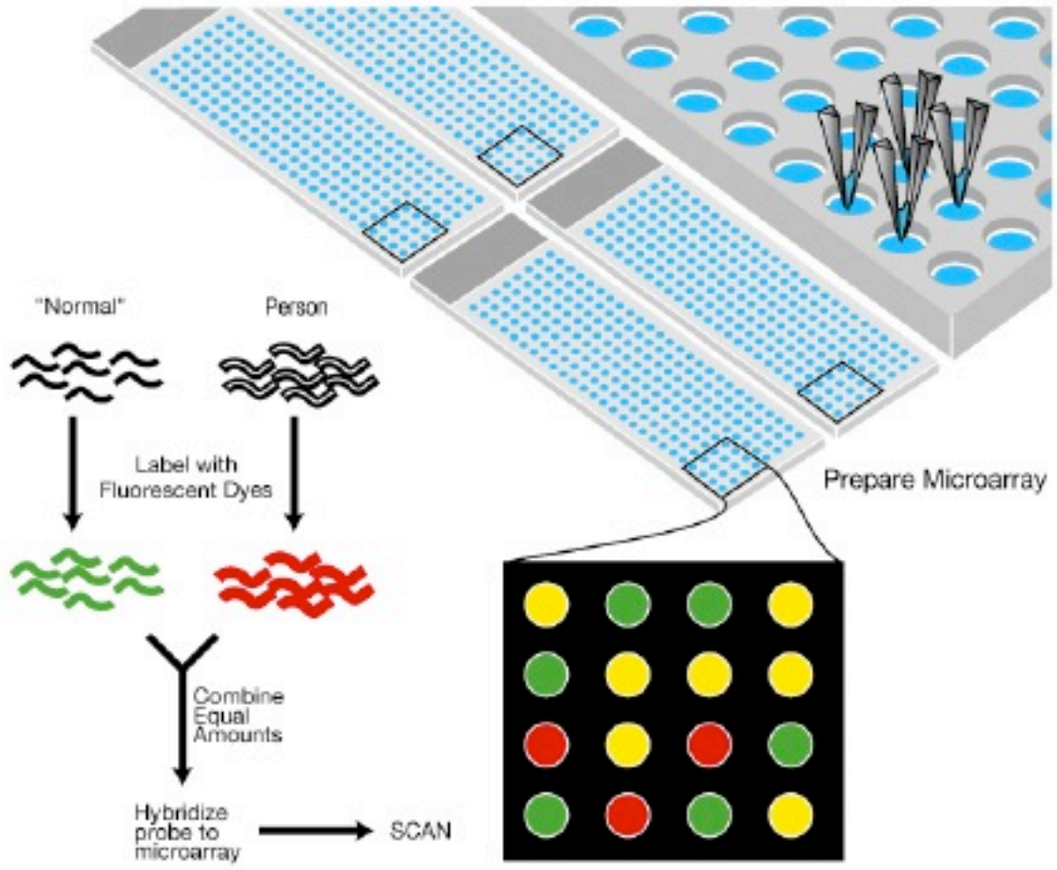
Segmental Dups

- track at UCSC Genome Browser (http://genome.ucsc.edu)
- Duplications of $>$ 1000 Bases of Non-RepeatMasked Sequence

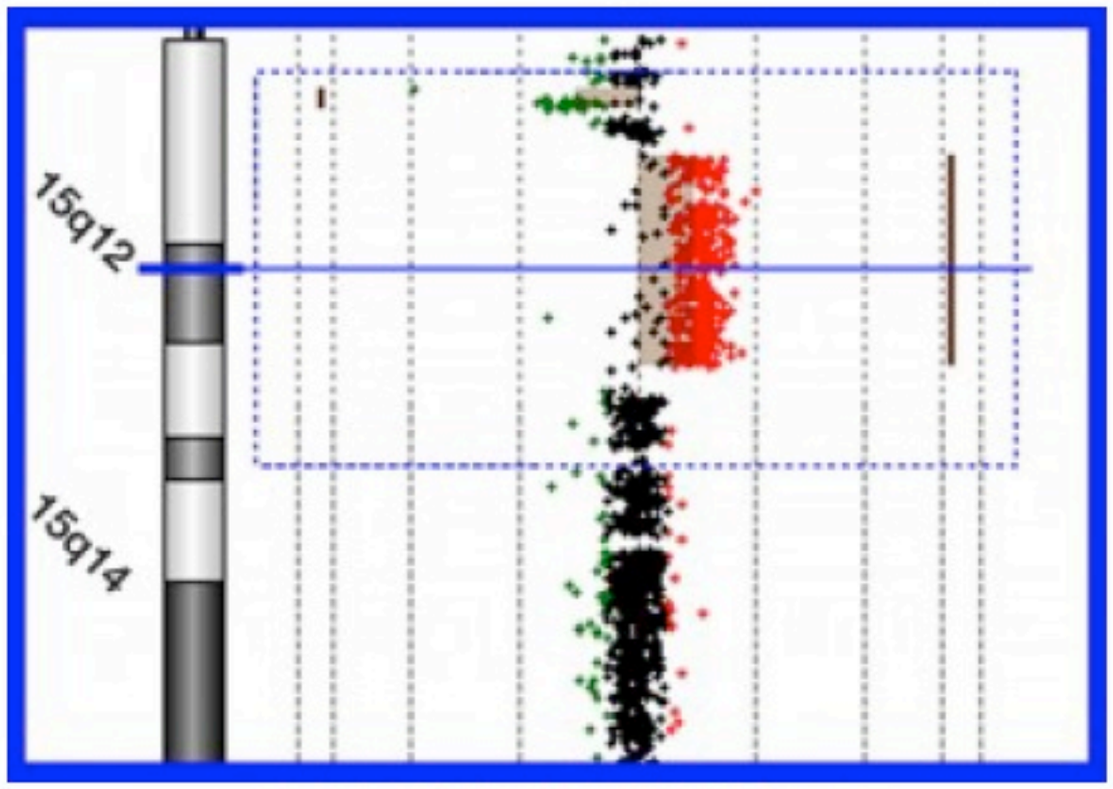


Chromosomal microarray analysis (CMA)

Over 25,000 patients from Baylor College of Medicine CMA database



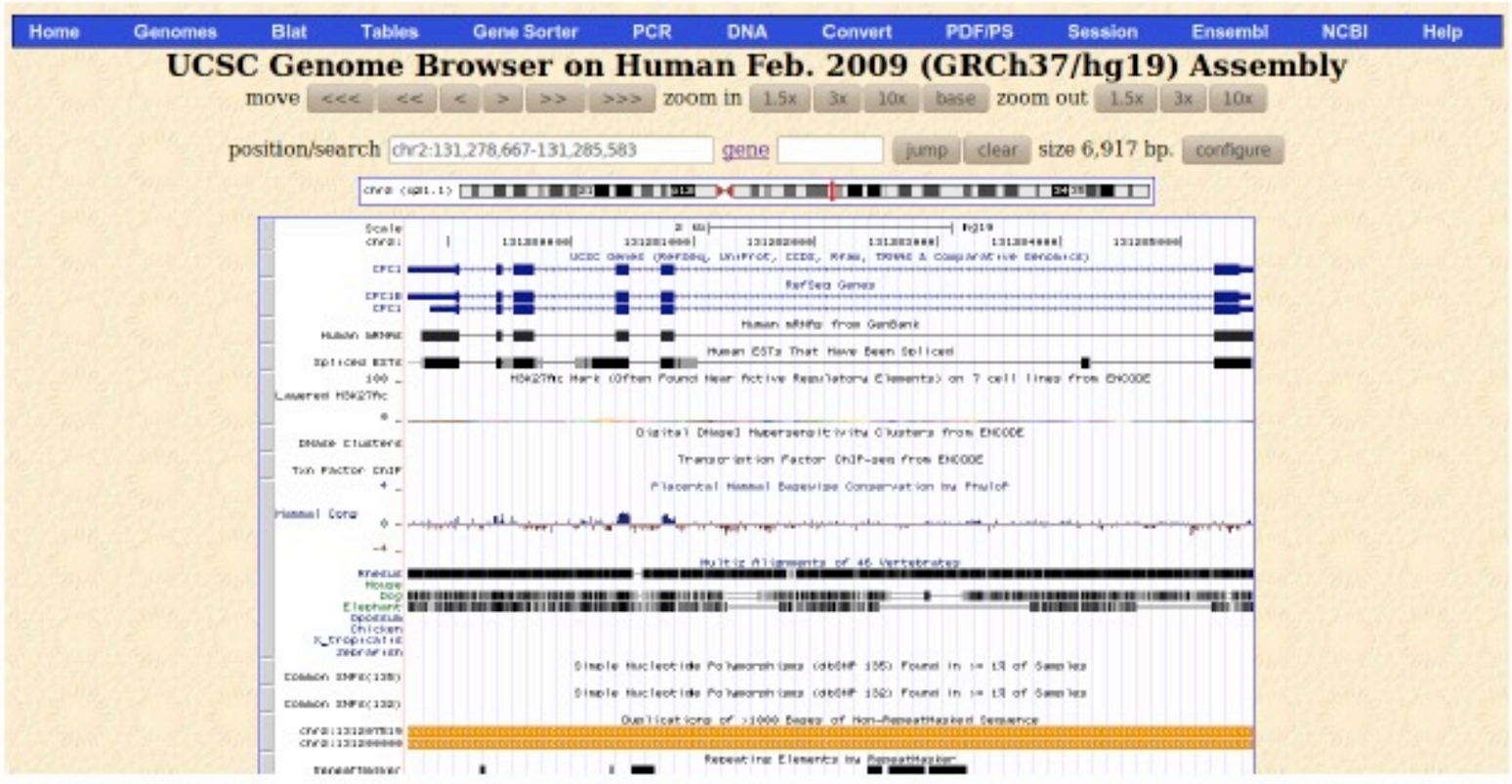
source: atlantichealth.dnadirect.com

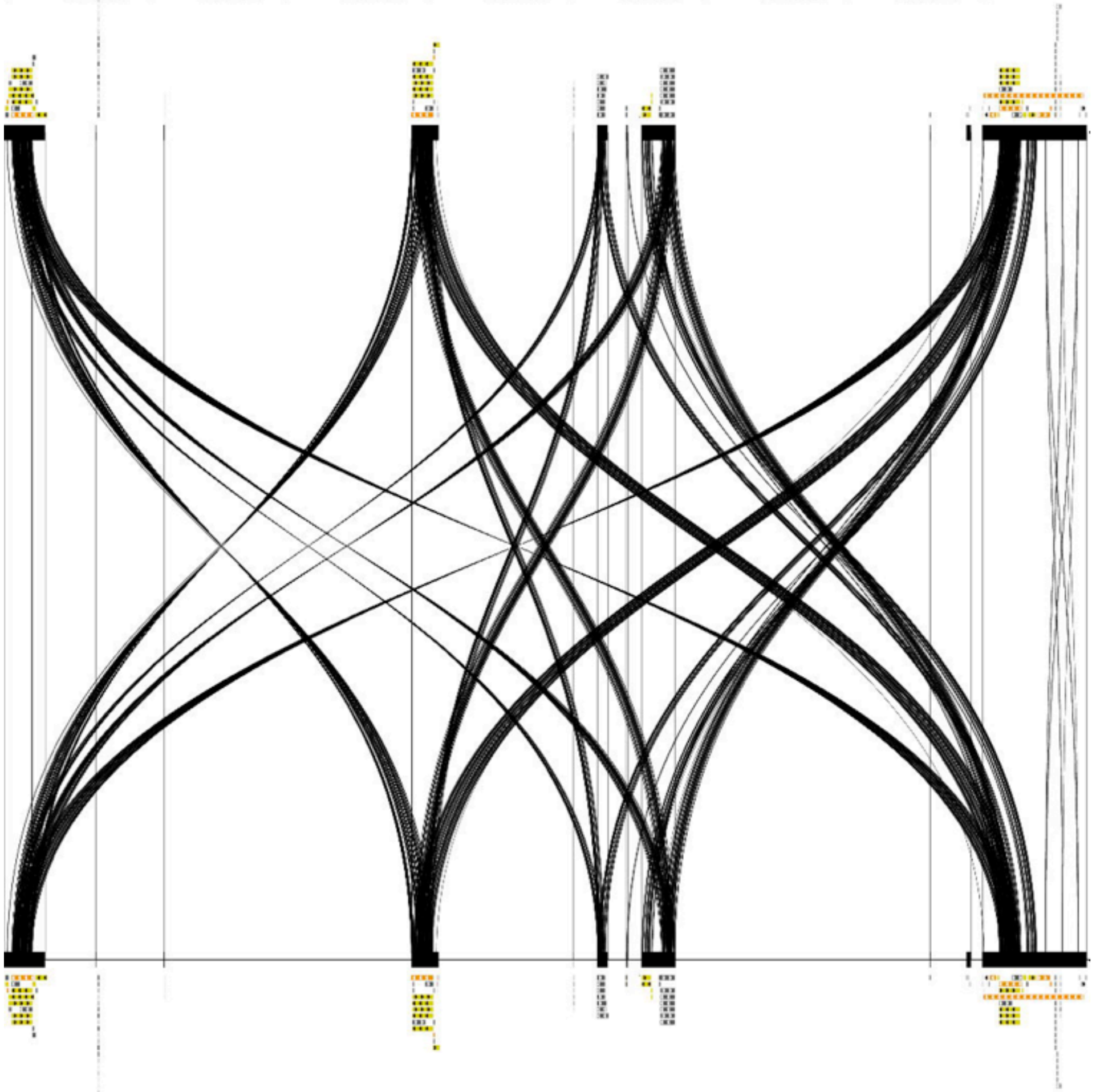


source: childrenshospitalblog.org

DP-LCRs

- direct paralogous LCRs (from SegDups)
- fraction matching (fractMatch) measure between paralogous elements at least 95%,
- rearrangement length: 50kb-10Mb
- not spanning centromeres
- minimal length of an element: 8 kb





1 Testowanie hipotez statystycznych

Testowanie hipotez pozwala nam z wykorzystaniem narzędzi statystyki testować hipotezy dotyczące analizowanych danych molekularnych. Możemy np szukać odpowiedzi na pytania w rodzaju:

1. Czy średnia ekspresja danego genu u pacjentów cierpiących na białaczkę typu AML jest różna od średniej ekspresji tego samego genu u grupy chorych na białaczkę typy ALL ?
2. Czy średnia ekspresja badanego genu jest niezerowa ?
3. Do jakiego stopnia ma rozkład normalny ?
4. Czy w badanej próbce znajdują się elementy odstające (*ang. outliers*)?
5. W jaki sposób wybrać cechy najlepiej dyskryminujące dwie badane populacje ?
6. Jal zbadać, czy częstości występowania danego motywu w różnych sekwencjach DNA jest taka sama?

1. Sformułuj **hipotezę zerową** H_0 oraz **hipotezę alternatywną** H_1. Najczęściej hipoteza zerowa odpowiada sytuacji nieciekawej, średniej, nie wyróżniającej się cechy. Natomiast odrzucenie hipotezy zerowej, równoznaczne przyjęciu hipotezy alternatywnej sugeruje, że rozważana cecha w istotny sposób dyskryminuje dwie populacje. Ponieważ decyzję o przyjęciu lub odrzucecniu hipotezy podejmujemy na podstawie danych, które traktujemy jako próbę losową, czyli realizację pewnego procesu losowego mamy niezerową szansę pomyłki. Odrzucenie poprawnej hipotezy zerowej określamy jako **błąd typu I**, natomiast przyjęcie fałszywej hipotezy zerowej nazywamy **błędem typu II**. Łatwo dostrzec, że obydwie te wielkości są ze sobą powiązane i dlatego musimy wybrać, którą z nich chcemy kontrolować. W zastosowaniach najczęściej konsekwencję różnych typów błędów są niesymetryczne i zazwyczaj przyjmuje się, że interesuje nas utrzymanie błędów typu I na odpowiednio niskim poziomie α (np. $\alpha = 1\%$ lub $\alpha = 5\%$).

2. Ustal poziom α dla błędów typu I.

3. Sformułuj odpowienią statystykę testową, czyli obliczaną na podstawie danych wielkość, której wartość będzie odpowiadała za przyjęcie bądź odrzucenie hipotezy zerowej. Jest to bardzo ważny krok i łatwo się zgodzić, że wybór kiepskiej statystyki zaważy na jakości testu.

4. Określ, które wartości statystyki testowej prowadzą do odrzucenia hipotezy zerowej. Wybór tych wartości jest taki, aby kontrolować poziom błędów α założony w kroku 2. W tym celu przydatne okazuje się pojęcie **p-wartości** (*ang. p-value*). Dla danej wartości statystyki testowej p-wartość jest zdefiniowana, jako prawdopodobieństwo uzyskania tej lub bardziej ekstremalnej wartości przy założeniu hipotezy zerowej. Jeśli tak policzona p-wartość jest mniejsza niż zakładany poziom błędów α, hipoteza zerowa zostaje odrzucona.

5. W ostatnim kroku analizujemy dostępne dane i sprawdzamy, czy wartość statystyki testowej odpowiada p-wartości pozwalającej nam odrzucić hipotezę zerową.

Z test

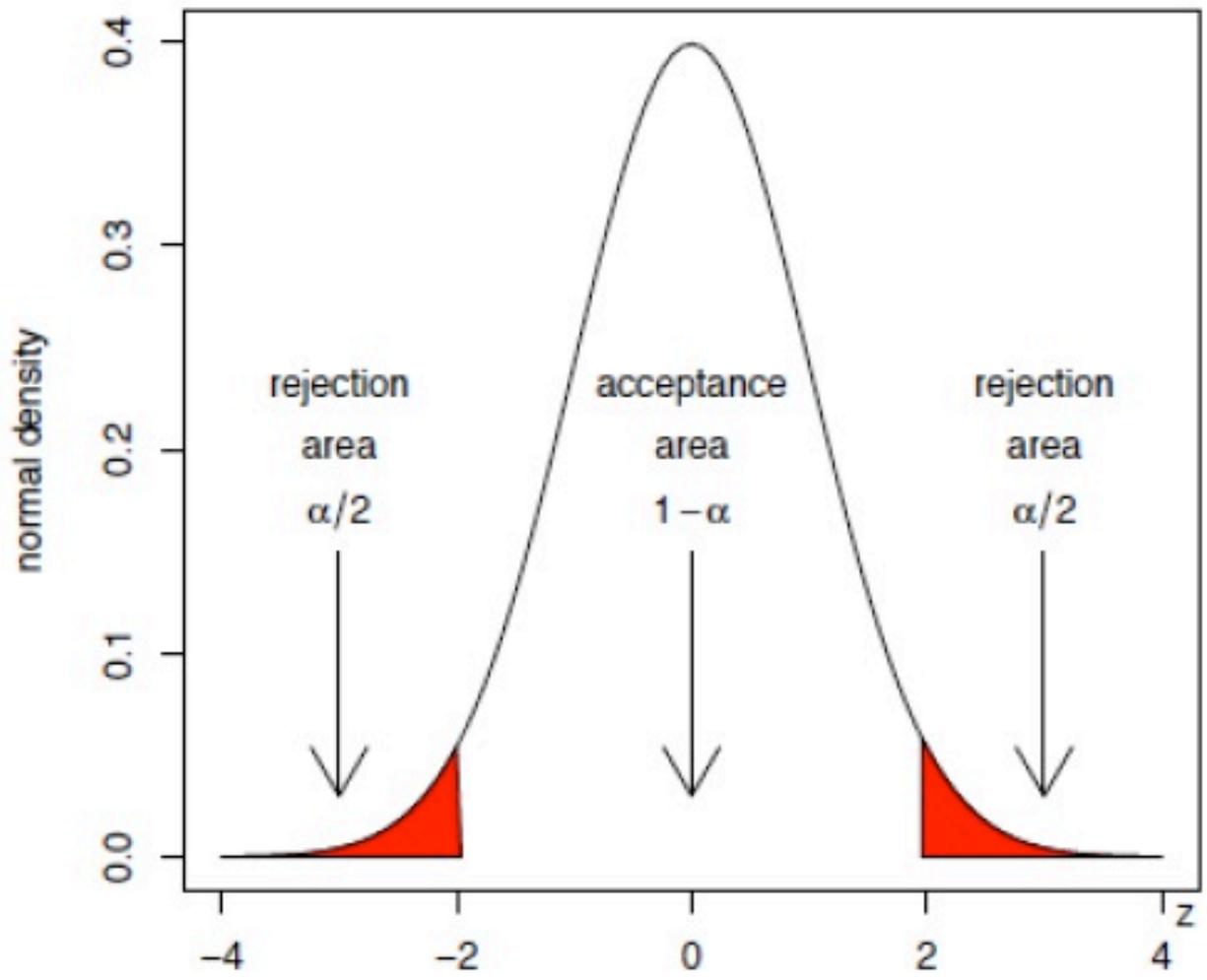
Zakładamy, że badane przez nas obserwacje $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego o nieznanej średniej i znanej wariancji σ. Jako hipotezę zerową przyjmujemy, że średnia wynosi μ_0, natomiast hipoteza alternatywna H_1 stwierdza, że $\mu \neq \mu_0$ (hipoteza złożona) lub $\mu > \mu_0$ (hipoteza prosta).

Liczymy następnie statystykę testową:

$$Z = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma}$$

Zmienna losowa Z ma standardowy rozkład normalny, a więc potrafimy policzyć P-wartość, czyli prawdopodobieństwo, że zmienna$Z$ przyjmie wartość bardziej ekstremalną niż $|z|$ – porównaj rysunek 1. Odrzucamy hipotezę zerową, jeżeli:

$$P(Z \leq -|z|) + P(Z \geq |z|) = 2P(Z \leq -|z|) \leq \alpha = 0.05$$



Rysunek 1: Obszar krytyczny dla testu Z.

jednopróbkowy T-test

Jeżeli wariancja badanej populacji nie jest znana, to do przetestowania hipotezy dotyczącej średnich używamy jednopróbkowego T-testu.

Niech $H_0 : \mu = \mu_0$ oraz $H_1 : \mu \neq \mu_0$. Statystyka testowa jest zmienną losową:

$$T = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

gdzie s^2 jest wariancją z próby, czyli:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

.

Ponieważ statystyka testowa ma rozkład T-Studenta o $n-1$ stopniach swobody potrafim policzyć P-wartość testu:

$$\text{P-val} = 2P(T_{n-1} \leq -|t|)$$

Odrzucamy H_0 jeśli jest ona mniejsza niż ustalony poziom istotność testu α.

dwupróbkowy T-test (Welch'a)

Załóżmy, że podobnie jak w dwóch poprzednich sytuacjach obserwacje pochodzą z rozkładu normalnego, tylko dysponujemy dwiema populacjami (np. próbki pobrane od osób zdrowych i chorych). Testujemy hipotezę $H_0 : \mu_x = \mu_y$ przeciwko $H_1 : \mu_x \neq \mu_y$. Załóżmy, że obserwacje w grupach wynoszą odpowiednio: $x_1, x_2, \ldots x_n$ oraz $y_1, y_2, \ldots y_m$. Niech $\bar{x}$ będzie

średnią dla pierwszej grupy, a $\bar{y}$ dla drugiej. Podobnie oznaczmy przez s_x^2 oraz s_y^2 wariancje z próby. T-statystykę obliczamy następująco:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Przykład 1 Hipoteza o zróżnicowanej ekspresji: *Załóżmy, że porównujemy poziom ekspresji pewnego genu w dwóch populacjach komórek (np. pochodzących od m zdrowych i n chorych dawców). Niech zmienne $X_1, X_2, \ldots, X_n$ oznaczają poziom ekspresji w zdrowych komórkach natomiast $Y_1, Y_2, \ldots, Y_m$ opisują populację chorych komórek. Zakładamy, że pomiary dotyczące poziomów ekspresji są niezależne i pochodzą z rozkładu normalnego o nieznanej wariancji σ^2 identycznej w obydwu grupach oraz nieznanych wartościach oczekiwanych μ_x oraz μ_y. Naturalna hipoteza zerowa mówi, że obydwie rozważane wartości oczekiwane są idenyczne, czyli $\mu_x = \mu_y = \mu$. Natomiast hipoteza alternatywna mówi, że są różne $\mu_x \neq \mu_y$, czyli badany gen w istotny sposób różnicuje dwie populacje. Okazuje, się że adekwatną statystyką w tym zadaniu jest statystyka t omówiona powyżej.*

dwupróbkowy T-test (przypadek równych wariancji)

Powróćmy do poprzedniej sytuacji, kiedy testujemy równość średnich w dwóch populacjach pochodzących z rozkładu normalnego przy założeniu, że wariancje obydwu rozkładów są równe: $\sigma_x^2 = \sigma_y^2$. Podobnie jak poprzednio testujemy hipotezę $H_0 : \mu_x = \mu_y$ przeciwko $H_1 : \mu_x \neq \mu_y$. Zdefiniujmy **łączną wariancję** jako:

$$s_{xy}^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Wtedy następująca zmienna losowa ma rozkład T-Studenta o $m+n-2$ stopniach swobody:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_{xy}^2 \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

F-test

Do tej pory zajmowaliśmy się testowaniem hipotez dotyczących średniej, czasem zakładaliśmy też (tak jak w poprzednim teście) że wariancje w dwóch badanych populacjach są równe.

F-test testuje tą właśnie własność: hipoteza zerowa zakłada, że $\sigma_x^2 = \sigma_y^2$, natomiast hipoteza alternatywna $H_1 : \sigma_x^2 \neq \sigma_y^2$. Statystyka testowa jest równa:

$$f = \frac{s_x^2}{s_y^2}$$

i ma rozkład F o $(n-1, m-1)$ stopniach swobody. s_x^2 oraz s_y^2 oznaczają wariancje z próby. Nie odrzucimy hipotezy zerowej jeśli:

$$P(F_{n-1,m-1} < f) \geq \frac{\alpha}{2} \text{ dla } f < 1 \text{ lub } P(F_{n-1,m-1} > f) \geq \frac{\alpha}{2} \text{ dla } f > 1$$

Test dwumianowy

Załóżmy, że badamy sekwencję mikroRNA i sformułowaliśmy hipotezę zerową mówiącą, że prawdopodobieństwo występowania puryny na danej pozycji w sekwencji jest równe $p = p_0$. Hipoteza alternatywna postuluje, że to prawdopodobieństwo jest większe $H_1 : p > p_0$. Po zsekwencjonowaniu sekwencji długości n okazało się, że występuje w niej k puryn. Zakładając rozkład dwumianowy dla H_0, możemy policzyć P-wartość, czyli prawdopodobieństwo zaobserwowania k lub więcej puryn przy założeniu hipotezy zerowej:

$$\text{P-val} = P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i}$$

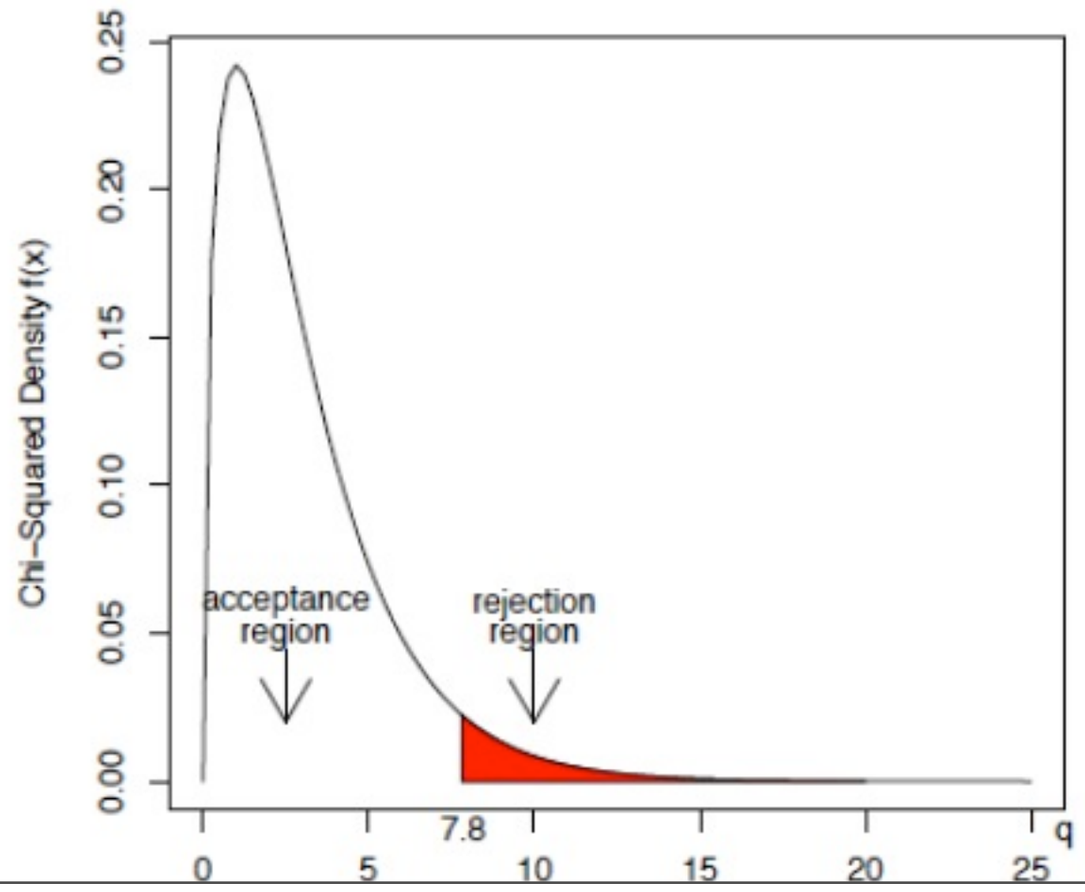
. Jeżeli dla otrzymanych danych k, n i przyjętego p_0, P-wartość jest odpowiednio mała odrzucamy hipotezę zerową. Bardziej ambitny przykład testu dwumianowego był wspomniany przy okazji rozkładu dwumianowego i dotyczył badania wzbogacenia adnotacji genów bliskich badanym obszarom genomowym.

Test chi kwadrat (Pearson)

Będziemy teraz testować hipotezę, która dotyczy więcej niż jednego parametru rozkładu, np niech $H_0 : (\pi_1, \pi_2 \ldots \pi_n) = (p_1, p_2, \ldots p_n)$ oraz $H_1 : (\pi_1, \pi_2 \ldots \pi_n) \neq (p_1, p_2, \ldots p_n)$. Jeśli badane parametry opisują prawdopodobieństwo uzyskania obserwacji danego rodzaju, to możemy policzyć oczekiwaną liczbę obserwacji i-tego rodzaju jako $e_i = np_i$, gdzie n jest rozmiarem badanej próbki. Niech o_i oznacza zaobserwowaną w próbce liczbę wyników i-tego rodzaju. Statystyka:

$$q = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

ma rozkład chi kwadrat o $n - 1$ stopniach swobody. Obszar krytyczny dla testu chi kwadrat o 3 stopniach swobody jest zilustrowany na rysunku 2.



Przykład 2 *Jako przykład zastosowania testu Pearsona rozważmy białko Zyksynę (składnik macierzy pozakomórkowej) i wysuńmy hipotezę zerową, że nukleotydy występują w tym białku z równymi częstościami:* $H_0 : (\pi_1, \pi_2, \pi_3, \pi_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. *Zyksyna składa się z* $n = 2166$ *nukleotydów, i przy założeniu hipotezy zerowej* $e_i = 541,5$ *dla* $i = 1, 2, 3, 4$. *Jeśli policzymy statystykę* q *dla* $o_i \in \{410, 789, 573, 394\}$, *przekonamy się, że* $q \approx 187$, *co odpowiada P-wartości: P-val*≈ 0, *czyli możemy z czystym sumieniem odrzucić hipotezę zerową.*

Test asocjacyjny chi kwadrat

Często dane, które badamy mają postać tabeli, której każda komórka jest zmienną losową:

	1	2	3	...	c	$\sum$
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1c}	$y_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2c}	$y_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	Y_{r1}	Y_{r2}	Y_{r3}	...	Y_{rc}	$y_{r.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	...	$y_{.c}$	y

Dla wyrobienia intuicji załóżmy, że powyższa tabela posiada jedynie dwa wiersze i dwie kolumny. Wiersze odpowiadają podziałowi według płci, a kolumny według ręczności (praworęczność vs leworęczność). Hipoteza zerowa zakłada, że nie istnieje żadna zależność pomiędzy płcią a leworęcznością, czyli prawdopodobieństwo spotkania leworęcznego mężczyzny jest takie samo jak spotkania leworęcznej kobiety.

Przy założeniu, że hipoteza zerowa jest poprawna, czyli kategorie wierszy i kolumn są od siebie niezależne, możemy obliczyć oczekiwaną liczbę obserwacji w komórce (j, k), czyli wartość oczekiwaną zmiennej losowej Y_{jk}:

$$E_{jk} = E(Y_{jk}) = \frac{y_{j.} y_{.k}}{y}$$

Jeśli hipoteza zerowa jest prawdziwa, to obserwowane wartości zmiennych Y_{jk} powinny być bliskie oczekiwanym, czyli znowu liczymy statystykę chi-kwadrat:

$$\sum_{jk} \frac{(Y_{jk} - E_{jk})^2}{E_{jk}}$$

która ma asymptotycznie rozkład chi kwadrat o $\nu = (r-1)(c-1)$ stopniach swobody.

Często sumy komórek w poszczególnych wierszach i kolumnach mogą być ustalone przed etapem testowania hipotezy (dlatego są oznaczane małymi literkami, jako że nie odpowiadają zmiennym losowym). Dodatkowo, żeby opisywany test był poprawny musimy założyć, że wszystkie obserwacje, które odpowiadają zmiennym zliczającym w komórkach tabeli są od siebie niezależne. W przypadku badania leworęczności dwóch bliźniaków jednojajowych ten warunek nie będzie spełniony. Z tego powodu testy asocjacyjne dla sekwencji DNA powinny być używane bardzo ostrożnie, ponieważ często rozważane organizmy są blisko spokrewnione i badane obserwacje nie są niezależne.

Przykład 3 *Jako przykład rozważmy test statystyczny na niezależność Markowa, który sprowadzi się do testu asocjacyjnego w tabeli* $4\ x\ 4$ *i będzie badał, czy częstość występowania danego nukleotydu na danej pozycji zależy od rodzaju nukleotydu na pozycji sąsiedniej.*

Przeprowadza się taki test, żeby ocenić czy można modelować sekwencję DNA jako ciąg prób Bernoulliego, czy raczej z użyciem łańcucha Markowa, który uwzględni taką zależność. Tabela asocjacyjna wygląda w naszym przypadku następująco:

	a	c	g	t	$\sum$
a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$y_{1.}$
c	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$y_{2.}$
g	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$y_{3.}$
t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	$y_{4.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{.4}$	y

Kategoria wiersza określa nukleotyd na pozycji i-tej, kategoria kolumny nukleotyd na pozycji $(i+1)$-szej, np: Zmienna losowa Y_{11} zlicza występowanie dinukleotydów aa w badanej sekwencji DNA. Hipoteza zerowa o niezależności Markowa odpowiada hipotezie zerowej o niezależności wierszy i kolumn. Dla większości sekwencji DNA będziemy zmuszeni odrzucić hipotezę zerową, ponieważ łańcuch Markowa lepiej modeluje sekwencję. Okazuje się też, że jeszcze lepiej modelują łańcuchy Markowa wyższego rzędu, a w niektórych przypadkach nawet niehomogeniczne łańcuchy Markowa.

Test dokładny Fishera

Test dokładny Fishera stosujemy do tablic wymiaru 2 x 2, w których komórkach zmienne losowe przyjmują niewielkie wartości. Dla dużych próbek możemy stosować omówiony powyżej test chi kwadrat, ale jeśli liczby w tablicy są mniejsze bądź równe 5, to postępujemy odmiennie.

	K	M	$\sum$
dieta	$a = 9$	$b = 1$	$a + b = 10$
brak diety	$c = 3$	$d = 11$	$c + d = 14$
$\sum$	$a + c = 12$	$b + d = 12$	$n = 24$

Załóżmy, że znamy wartości brzegowe, czyli $a+b$, $c+d$, $a+c$ oraz $b+d$ w powyższej tabeli. Dodatkowo przyjmijmy, że przedstawione liczby zliczają osoby stosujące i nie stosujące dietę w losowej próbie 12 kobiet i 12 mężczyzn. Hipoteza zerowa stwierdza, że płeć i decyzja o stosowaniu diety są zmiennymi niezależnymi, czyli kobiety tak samo często jak mężczyźni przechodzą na dietę. Przy założeniu hipotezy zerowej, P-wartość dla naszej tabeli możemy policzyć stosując rozkład hipergeometryczny:

$$\text{P-val} = P(Y_{11} = a, Y_{12} = b, Y_{21} = c, Y_{22} = d) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Test Shapiro-Wilka

Załóżmy, że chcemy sprawdzić, czy nasze obserwacje: $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego. W wielu metodach statystycznej analizy danych czyni się takowe założenie, dlatego bardzo ważne jest sprawdzenie czy jest ono prawdziwe, albo chociaż, czy nie będziemy zmuszeni odrzucić hipotezy zerowej mówiącej o normalności rozkładu. Popularnym testem adekwatnym w tym przypadku jest **test Shapiro-Wilka** o następującej statystyce:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie $x_{(I)}$ jest i-tym co do wielkości elementem spośród $x_1, x_2, \ldots x_n$, natomiast stałe a_i dla $i = 1, \ldots n$ wynoszą:

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

Wektor $m = (m_1, m_2, \ldots m_n)^T$ odpowiada wartościom oczekiwanym **statystyk porządkowych**[1] dla niezależnych zmiennych losowych o standardowym rozkładzie normalnym, a macierz V jest ich macierzą kowariancji.

Omówimy teraz testy pozwalające sprawdzić, czy wśród badanych obserwacji są elementy odstające. Hipoteza zerowa twierdzi, że takowych nie ma, natomiast hipoteza alternatywna stawia tezę, że wśród naszych danych jest co najmniej jeden taki odstający element. Przy założeniu, że dane $X_1, X_2, \ldots X_n$ pochodzą z rozkładu normalnego (warto to przetestować), hipotezę tą testuje **test Grubbsa**. Statystyka testowa jest równa:

$$G = \frac{\max_{i=1\ldots n} |X_i - \bar{X}|}{s}$$

Odrzucimy hipotezę zerową na poziomie istotności α jeżeli

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{n-2}(\frac{\alpha}{2n})}{n-2+t^2_{n-2}(\frac{\alpha}{2n})}}$$

gdzie $t^2_{n-2}(\frac{\alpha}{2n})$ jest górną wartością krytyczną w rozkładzie T o $n-2$ stopniach swobody, przy poziomie istotności $\frac{\alpha}{2n}$. Czyli taką wartością, że całka z górnego ogona tego rozkładu liczonego od tej wartości wynosi $\frac{\alpha}{2n}$.

Test Manna-Whitneya

Załóżmy, że obserwowane wartości zmiennych losowych $X_1, \ldots X_n$ oraz $Y_1, \ldots Y_m$, czyli $x_1, \ldots x_n$ oraz $y_1, \ldots y_m$ są podane w kolejności rosnącej. Każdej obserwacji przypisujemy jej pozycję rankingową (*rangę*) na wspólnej posortowanej liście, czyli jedną z liczb $1, 2, \ldots m + n$. Oczywiście jest to możliwe tylko wtedy, gdy wszystkie obserwacje są różne co niniejszym dla uproszczenia założymy (oczywiście istnieje też ogólna postać tego testu dopuszczająca powtórzenia). Statystyką testową jest suma *rang* obserwacji z pierwszej grupy. Łatwo policzyć, że suma rang wszystkich obserwacji (z obydwu grup) wynosi $R = (m+n)(m+n+1)/2$. Przy założeniu hipotezy zerowej (rozkłady z jakich pochodzą obserwacje są takie same) suma rang obserwacji w pierwszej grupie powinna stanowić $n/(m + n)$ część sumy wszystkich rankg R, czyli jej wartość oczekiwana wynosi:

$$\frac{n}{n+m}\frac{(n+m)(n+m+1)}{2} = \frac{n(n+m+1)}{2}$$

Można też pokazać, że suma rang w pierwszej grupie ma rozkład bliski normalnemu, natomiast wariancja tej sumy wynosi:

$$\frac{nm(n+m+1)}{12}$$

Podsumowując zredukowaliśmy zadanie do testowania wartości średniej rozkładu normalnego o znanej wariancji, co już potrafimy zrobić.

Test permutacyjny

Przy założeniu hipotezy zerowej (dane w próbkach pochodzą z jednakowych rozkładów), wszystkie $\binom{n+m}{n}$ permutacje przypisujące pierwsze n elementów do pierwszej grupy i kolejne m elementów do drugiej grupy są jednakowo prawdopodobne. Dla każdej takiej permutacji liczymy wartość pewnej statystyki testowej (nie jest jednoznacznie powiedziane jaką wybrać może to być np. statystyka dla dwupróbkowego T testu). Jedna z wartości statystyki będzie tą, która pojawia się dla permutacji prawdziwych obserwowanych danych. Jeśli hipoteza alternatywna twierdzi, że średnia z pierwszej grupy jest większa od średniej z drugiej grupy, to używając parametru α (poziom błędów typu I) odrzucamy hipotezę zerową jeśli obserwowana wartość statystyki jest pośród górnych $\alpha 100\%$ wartości.

Zauważmy, że przy tym podejściu nie potrzebujemy znać rozkładu prawdopodobieństwa badanej statystyki testowej. Dodatkowo jeśli używamy T statystyki nie musimy jej obliczać dla każdej permutacji. Wystarczy jedynie policzyć różnicę średnich w dwóch grupach dla każdej permutacji, a tak naprawdę to wystarczy tylko policzyć średnią obserwacji w pierwszej grupie[2]. Ostatnią miłą własnością testu permutacyjnego jest fakt, że dla danych pochodzących z rozkładu normalnego otrzymane wyniki są bardzo zbliżone do tych z T testu.

Niemiłą z kolei własnością jest złożoność obliczeniowa tego testu (nawet dla średnich wartość n i m liczba permutacji rośnie bardzo szybko). W sytuacji kiedy nie możemy policzyć statystyki testowej dla wszystkich permutacji, losujemy odpowiednio dużą próbkę permutacji i odrzucamy hipotezę zerową jeśli obserwowana wartość statystyki testowej znajdzie się pośród $\alpha 100\%$ rozważanych dodatnich wartości.

8 Jednoczesne testowanie wielu hipotez

Wróćmy teraz do naszych wielowymiarowych danych. Filtrując cechy np. przy pomocy testu t chcielibyśmy zredukować wymiar danych poprzez wybór cech które w statystycznie istotny sposób różnicują dwie populacje. Zauważmy, że nasze zadanie jest równoważne problemowi jednoczesnego testowania wielu tysięcy hipotez zerowych: $H_1, H_2, \ldots, H_m$. Oznaczmy przez R liczbę odrzuconych hipotez (czyli np. w przypadku mikromacierzy liczbę genów o różnicującej ekspresji). Mamy następującą sytuację:

	# przyjętych H_0	# odrzuconych H_0	$\sum$
# prawdziwych H_0	U	V	m_0
# fałszywych H_0	T	S	m_1
$\sum$	$m - R$	R	m

gdzie R jest obserwowaną zmienną losową, m_0 oraz m_1 nieznanymi parametrami, podobnie jak U, V, T oraz S są nieobserwowanymi zmiennymi losowymi.

Opiszemy teraz jak uogólnia się zadanie kontroli błędów typu I w problemie testowania wielu hipotez. W przypadku pojedynczego testu (hipotezę zerową oznaczamy tutaj dość myląco przez H_1) potrafiliśmy policzyć wielkość c_α, taką że:

$$\Pr(|T_1| \geq c_\alpha | H_1) \leq \alpha$$

gdzie T_1 jest wartością statystyki testowej. Odrzucaliśmy hipotezę H_1 jesli $|T_1| \geq c_\alpha$. Najczęściej stosowane uogólnienia tego podejścia są następujące (por. [1]):

- **PCER** (*ang. Per-comparison error rate*), miara jest zdefiniowana jako średnia z wartości oczekiwanej błędów typu I, czyli:

$$\text{PCER} = \frac{E(V)}{m}$$

- **PFER** (*ang. Per-family error rate*), odpowiada oczekiwanej liczbie błędów typu I:

$$\mathrm{PFER} = E(V)$$

- **FWER** (*ang. Family-wise error rate*), jest zdefiniowana jako prawdopodobieństwo co najmniej jednego błędu typu I:

$$\mathrm{FWER} = \Pr(V \geq 1)$$

- **FDR** (*ang. False discovery rate*), definiujemy jako oczekiwaną proporcję błędów typu I pomiędzy odrzuconymi hipotezami zerowymi (jest to procent fałszywych pozytywów, czyli cech uznanych niesłusznie za istotne):

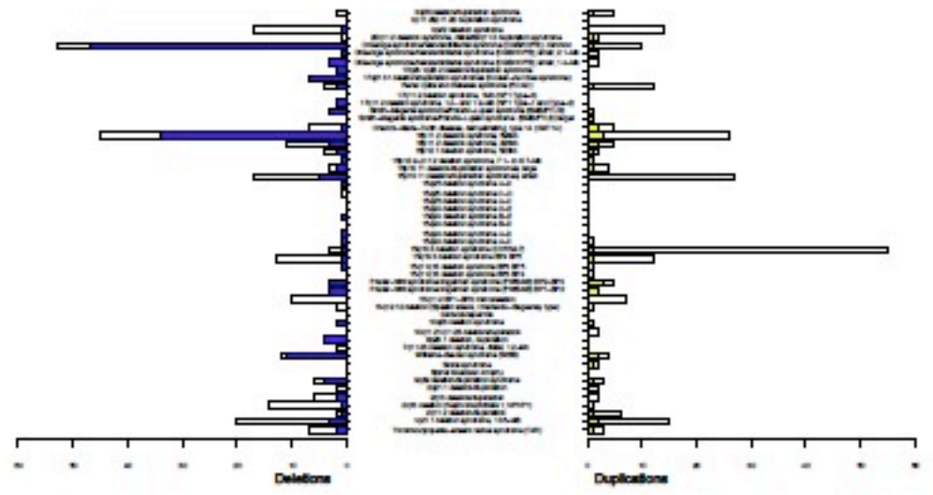
$$\mathrm{FDR} = E(Q)$$

$$Q = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}$$

Omówmy bardziej szczegółowo tylko dwie ostatnie miary jako najczęściej stosowane w bioinformatyce. W przypadku FWER stosujemy tzw **poprawkę Bonferroniego** i odrzucamy hipotezę zerową H_j $(j = 1, 2, \ldots, m)$ jeśli odpowiednia p-wartość jest mniejsza bądź równa $\frac{\alpha}{m}$ (gdzie α jest dopuszczalnym procentem błędów I typu w pojedynczym teście).

Genomic features correlating with NAHR frequency

- frequency for known pathogenic syndromes (de-novo deletions considered)
- sets of DP-LCRs pairs flanking these syndromes



Two steps of statistical analysis:

1. exploratory analysis with the Spearman's rank correlation
2. quasi-Poisson regression model

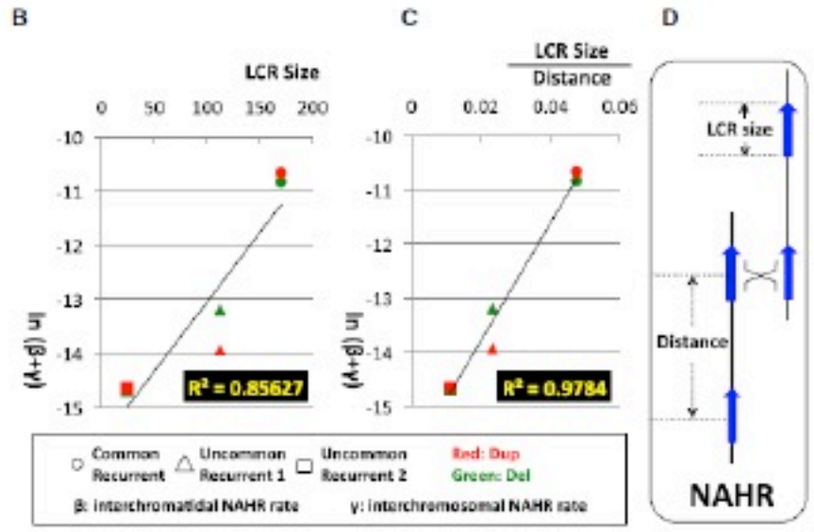
Considered parameters

DP-LCR:

- average lengths, distances, fraction matching
- presence of the 13-mer recombination hotspot motif 5'-CCNCCNTNNCCNC-3'

LCR clusters:

- number of LCRs within the cluster, average length of LCRs,
- concentration of recombination hotspot motif
- multiple statistics of these parameters (median, lower, and upper hinge, minimum and maximum)



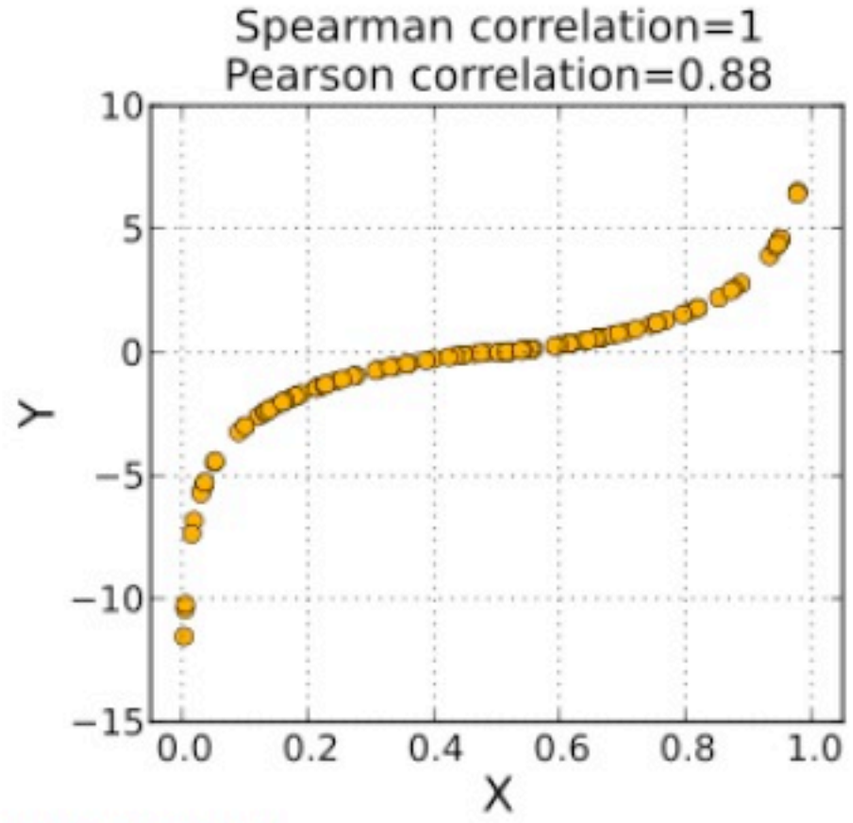
Liu et al. 2011

Spearman's rank correlation

- Pearson correlation coefficient between the ranked variables.
- assesses how well the relationship between two variables can be described using a monotonic function

Sample of size n, the n raw scores X_i, Y_i are converted to ranks x_i, y_i

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$



source:Wikipedia

Statistical modeling based on genome-wide studies and CMA data

On a genome-wide scale, we found that the following properties of DP-LCRs correlate with NAHR frequency:

- length of homology (weak association, Spearman corr., p=1.68e-01);
- distance between homologous pair; inverse relationship - the further the DP-LCR are apart, the less frequent (Spearman corr., p=2.19e-04);
- percent DNA sequence identity (i.e. fraction matching of DP-LCRs, p=8.18e-05).

Feature of LCR cluster	Comparison of LCR clusters flanking active NAHR hot spots vs. LCR clusters flanking inactive cold spots (*p*-values from Mann-Whitney-Wilcoxon test)		
	Feature is greater in LCR clusters flanking active NAHR hot spots	**Feature is greater in LCR clusters flanking inactive NAHR cold spots**	**Spearman rank correlation coefficients and *p*-values**
GC content within the cluster	*****(*p*=1.11e-04)**		**0.54** (*p*=7.04e-03)**
Minimum length of homology among LCRs within the cluster		(*p*=9.96e-01)	0.12 (*p*=5.74e-01)
First quartile of the length of homology among LCRs within the cluster		(*p*=8.43e-01)	0.02 (*p*=9.26e-01)
Median length of homology among LCRs within the cluster		(*p*=5.57e-01)	0.23 (*p*=2.71e-01)
Third quartile of the length of homology among LCRs within the cluster		(*p*=4.81e-01)	0.15 (*p*=4.73e-01)
Maximum length of homology among LCRs within the cluster	(*p*=1.41e-01)		**0.41* (*p*=4.62e-02)**
Total number of occurrences of the 13-mer recombination hot spot motif in the cluster		(*p*=2.7e-01)	**0.51* (*p*=1.17e-02)**
Minimum number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster		(*p*=2.25e-01)	0.00 (*p*=1.00)
First quartile of the number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster		(*p*=7.59e-01)	0.01 (*p*=9.33.e-01)
Median number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster	(*p*=7.1e-02)		**0.48* (*p*=2.01e-02)**

Collaboration

- Piotr Dittwald - University of Warsaw, Poland
- Paweł Stankiewicz, Tomasz Gambin, James Lupski, Sau Wai Cheung - Baylor College of Midicine, Houston, TX
- ...

See also:

Dittwald et al., NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits, Genome Research, 2013.