statisticon

STATISTICS & RESEARCH

Från ebola i smutsiga tält till NästGenerationsSekvensering för $miljoner$ – it’s bioinformatics!

FMS vårmöte 2015-03-17

Ingrid Lönnstedt

Statistikens roll inom bioinformatik

Igår – Idag - Imorgon

Vad är bioinformatik?

- Wikipedia: An interdisciplinary field that develops methods and software tools for understanding biological data.
- As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data.

- Dataanalys inom molekylär biologi/genetik/genomik/omics
- Ofta högdimensionella data i massor
  - Arraydata
  - Nästgenerationssekvensering (NGS)
  - …

Vem jobbar med bioinformatik?

Bioinformatics. Current Limitations and Insights for the Future

Plant Physiology **2005**, Seung Yon Rhee.

- Biologer kommer att
  - spendera mer tid vid datorn än i labbet.
  - behöva mer kontakter med datavetare, statistiker etc
- Ägandet av data kommer att förändras
  - Man kommer att analysera andras data
  - Det blir viktigare med hög kvalitet och struktur på data
- Problem:
  - Det saknas normer för hur data analyseras, annoteras och visas deskriptivt.

Läget idag? Statistik i dagens genetikforskning?

The DNA Exchange

HOME ABOUT US AUTHOR BIOS HOUSE RULES

← Questioning the "genetic counselor" professional title

Resta's Rules of Genetic Counseling →

BY LAURAHERCHER | DECEMBER 19, 2014 · 1:27 PM

↓ Jump to Comments

TOP TEN GENETICS STORIES FOR 2014: INCREDIBLE ACHIEVEMENTS (DON'T COME CHEAP)

No question that 2014 has been a year to celebrate for the field of genetics. Stem cell therapy, gene therapy, next-generation sequencing as a reliable clinical tool: we may not be there yet, but surely we are close. We have reached the suburbs of the Promised

Enter your email to receive notifications of new blog posts.

Join 754 other followers

Enter your email address

Get Email Updates

Twitter

- RT @GenesinLife: Is 'genetic counselor' the best name for the profession? Interesting article by @TheDNAExchange looks deeper. http://t.co/... 1 month ago
- RT @MishaAngrist: "How do you simplify a message, when the message itself is intended to express

10: Sekvenseringsteknik

Illumina HiSeq X

- 5+ genomes per day
- $1000 per genome
- (10 million dollars to buy equipment)

NGS: Next Generation Sequencing
Vad är det?

NästGenerationsSekvensering

En metod för att identifiera intressanta fragment ur DNA-kod:

DNA-fragment som

- påverkas av behandling
- som skiljer sig mellan sjuka och friska
- som är involverade i en viss funktion

Next Generation Sequencing data

Count segments overlapping (or starting at) each genomic position

```
…111111111123333333345555555552222211000000000…

                        TCTTATCTAGGAGC
                         CTTATTTAGGAGCTC
            CTTTCTAGATCTTATCTAG
             TTTCTAGATCTTATTTA
  TAGCTAGAGACTTTCTAGATCTTATC

…TAGCTAGAGACTTTCTAGATCTTATCTAGGAGCTCGCGCTAAGT…
```

Hitta intressanta regioner bland 3 miljarder genetiska positioner

Exempel från epigenetik:

Bakgrundsbrusnivå

Intressant region eller bara slumpmässig förhöjning i täckning?

9. Paralyzed man walks again

- After transplant of olfactory sheating cells grown in culture
  - With walker
  - No replication – but guys, he walks!

8: IQ bestäms av gener

Studie där 25’000 individer gjort test i kognitiv förmåga

Resultaten visar korrelation med 3 SNPs

Dessa SNPs uppskattas höja IQ med 1/3 steg vardera.
För att förklara 15% av variationen i IQ behövs mer än 1 miljon individer.

Single Nucleotide Polymorphism (SNP)

Allele A ...GGTAGCCCATGCAAGAGTTACTACAG...

Allele B ...GGTAGCCCATGCGAGAGTTACTACAG...

- A single nucleotide alteration in genomic DNA
- Every 100-300 bps of the genome
- Accounts for ~90% of genomic variation
- ~7 million SNPs, allele frequency > 5%
- Evolutionary stable

SNP arrays

- Measure whether a sample has
  - Standard allele (A), or
  - Alternative allel (B)

at each SNP position across the genome.

- Measures fluourescent dye on array for each SNP

GWAS

Sök korrelation mellan allel och IQ.

Kuriosa

From Business Insider:

BRIEFING

The Difference Between Being Smart Or Not Comes Down To Three Genes

CHRIS PASH

From an editorial at RealClearEducation:

September 11, 2014

Finally, Some Explanation on the Genetic Basis of Intelligence

By Daniel Willingham

science 2.0 JOIN THE REVOLUTION

Intelligence Inheritance – 3 Genes That Add To Your IQ Score

By The Conversation | September 8th 2014 07:30 PM | Print | E-mail | Track Comments

The Sydney Morning Herald

Technology

Digital Life | IT Pro | Sci-Tech | Other Tech News | Gadgets on the Go | Compare & Save

You are here: Home > Technology > Sci-Tech >

Scientists discover 'smart' genes

7. Sekvensering i kampen mot Ebola

Sekvensering av ebolavirus från 78 patienter i Sierra Leone

- Förbättrad diagnosmetod
- Kartläggning av virusets utveckling och ursprung.

Av 58 författare till Science-artikeln i september var 5 döda innan oktobers slut.

6. Friska celler återställer muterad funktion i möss

In mice with mutated FAH gene and human congenital liver disease, normally functioning genes were inserted and resulted in enzyme-producing liver cells.

- Measurable improvement occurred in the health of the mice.

5. Mosaikgenom

Hur stor andel av genomet delar jag med mitt barn?

- En mamma från Washington höll på att förnekas vårdnaden för de 3 söner hon fött
  - Hon visade sig ha två genom pga ett tivllingfoster hon absorberat när hon var själv var ett foster.
- 1/25 barn med nya mutationer (deletions) visade sig förklaras av mosaikgenom hos en förälder.
- 30% av kausala mutationer hos barn med hjärnskador har visat sig förklarade av mosaikgenom hos förälder.

Att beakta när vi söker efter genetiska orsaker eller informerar familjer om deras risk för ovanliga sjukdomar!

4. Stamcellsterapi

- Blinda patienter har återfått synen.
- Förbättring för strokepatienter.
- 40 patienter har fått implantat av pancreatic progenitor cells som mognar och börjar producera insulin
- …

3. Vem ska få behandling?

- 2012 godkändes Kalydeco som hjälper 4-5% av patienter med Cystisk Fibros.
  - $300,000 per år och patient
- 2014: Tre kvinnor i Arkansas stämmer Medicaid för att de inte får rätt till Kalydeco
  - De måste bevisa att konventionella läkemedel inte hjälper dem.
  - Etablerade läkemedel ger biverkningar (avsevärt fler infektioner och lungproblem som skadar patienterna permanent på sikt) men ger acceptabel lungfunktion enligt studier.

Fastställ genprofil med arraydata

Cellprov...

…spottas ut på en preparerad array...

…en prick för varje gen/SNP talar om hur stor aktivitet den genen har i provet/proverna…

…genom att varje prick får olika optisk intensitet.

Arraydata

Tusentals stickprov: en för varje gen

Gene	logratio1	logratio2	logratio3	mean	sd	t
AA676466	0.08098692	0.1102512	0.006119115	0.06578575	0.05370456	0.7072289
T60048	0.2266108	0.02714003	0.1947391	0.1494967	0.1071555	0.8054827
T60048	0.1943610	0.1384412	0.2147096	0.1825039	0.03949254	2.668066
AA481464	-0.1768410	-0.02858052	-0.06959967	-0.09167372	0.07655545	-0.6913661
AA447774	-0.07609968	-0.02876668	-0.1815102	-0.09545884	0.07819031	-0.7048595
AA427899	-0.1280369	-0.1406169	-0.2190607	-0.1625715	0.04932376	-1.902951
AA663995	-0.962603	0.02947815	-0.848488	-0.593871	0.5428431	-0.6316218
AA485911	0.1254507	0.2218707	0.01776646	0.1216959	0.1021039	0.6881342
H69583	0.3840236	1.870715	0.4454625	0.9000672	0.8411671	0.6177774
AA427899	-0.04957437	-0.009066125	-0.1368397	-0.06516008	0.06529708	-0.5761389
AA464600	1.173685	1.286068	0.7869012	1.082218	0.2618522	2.386151
AA670422	0.1724398	0.3050201	0.1418451	0.2064350	0.08673681	1.374103
AA454852	-0.1645123	-0.0002256093	-0.08410652	-0.08294816	0.08214949	-0.5829633
AA608558	-0.1786685	-0.02626047	-0.1013238	-0.1020843	0.07620686	-0.7733998
AA888182	-0.1496114	0.01798421	-0.1504318	-0.09401967	0.09699907	-0.5596165

2. Genterapi kommersiellt tillgänglig

- Glybera, a cure for lipoprotein lipase deficiency (rare disease which gives sky high triglycerides, eruptive fat-filled lesions, frequent abdominal pain and bouts of pancreatitis)
  - $1.4 million per patient : ONCE

1. Whole Exome database of 60,000+ genomes online

- Exome Aggregation Consortium
- Released in October
- Users crashed servers on day 1.
- 120,000 views from 17,000 users in month 1.
- Massor av data i en och samma viewer.

An example genome viewer

The IGV application window.

Thorvaldsdóttir H et al. Brief Bioinform 2013;14:178-192

© The Author(s) 2012. Published by Oxford University Press.

Briefings in Bioinformatics

Varje dataset kostar $miljoner$ och folk delar gärna med sig! Varför?

- Forskningsfrågorna kräver mer data än vad någon enskild forskningsenhet kan samla in själv.
- De genetiska fenomen forskningen hittar är ovanliga och kräver tillräcklig styrka och sofistikerade metoder.

Ett statistiskt problem som dessa databaser skapar:

Är de olika dataseten jämförbara?

RUV Normalisering av högdimensionella data

- Remove Unwanted Variation
- High-dimensional data (e.g. array data in growing databases)
- Negative control genes: Assume no change
- Positive control genes: Assume change

Justera varje array enligt ”principalkomponenter” hos negativa kontrollgener.

Använd fördelningen av p-värden för att välja bästa möjliga justering.

Gagnon-Bartsch, J.A. and Speed, T.P. Biostatistics. 2012

Bioinformatics. Current Limitations and Insights for the Future

Plant Physiology **2005**, Seung Yon Rhee.

- Biologer kommer att
  - spendera mer tid vid datorn än i labbet.
  - behöva mer kontakter med datavetare, statistiker etc
- Ägandet av data kommer att förändras
  - Man kommer att analysera andras data
  - Det blir viktigare med hög kvalitet och struktur på data
- Problem:
  - Det saknas normer för hur data analyseras, annoteras och visas deskriptivt.

Bioinformatik

- Bioinformatik erbjuder en mängd statistiska utmaningar – kul!
- I grunden handlar det om helt vanlig statistik, om än ofta applicerad på stora dataset.
- Sofistikerade statistikmetoder förbättrar forskningsresultat och sätter god standard – viktigt!
- Prova!

statisticon
STATISTICS & RESEARCH

Statisticon AB
Vi ser livet bakom siffrorna
+46 (0)10 130 80 00, info@statisticon.se, www.statisticon.se
Östra Ågatan 31, SE-753 22 Uppsala Klara Södra Kyrkogata 1, Box 837, SE-101 36 Stockholm Sweden