KØBENHAVNS UNIVERSITET

Fagligt samspil for
Bioteknologi & Matematik
på STX & HTX

Dataanalyse og kræft

Dataanalyse som grundlag for kræftforskning og udvikling af præcisionsmedicin

Bioteknologi

Baggrundsartikler

Matematik

Forord

Formålet med dette undervisningsmateriale er at give danske gymnasielever indblik i, hvordan genetiske data analyseres og anvendes i forbindelse med kræftforskning og udvikling af præcisionsmedicin. Materialet giver eleverne mulighed for at arbejde med emnerne kræftbiologi, bioinformatik og præcisionsmedicin samt statistik og matematiske modeller med udgangspunkt i autentiske genomiske datasæt fra open source databasen cBioPortal.

Målgruppen er gymnasieelever i 3g på htx og stx med Bioteknologi A og Matematik A.
Materialet er udarbejdet som et fagligt samspil mellem Bioteknologi og Matematik, men kan også anvendes som rent Bioteknologi- eller Biologimateriale.

Undervisningsmaterialet er udarbejdet i et projektsamarbejde mellem *Biotech Research and Innovation Centre* (BRIC) på Københavns Universitet, *Institut for Folkesundhedsvidenskab* på Københavns Universitet, Afdeling for Patologi på Herlev og Gentofte Hospital, Region Hovedstaden, samt Roskilde Gymnasium. Projektet er finansieret af Novo Nordisk Fondens pulje *Projektstøtte til naturvidenskabelig uddannelse og formidling*.

Materialet indeholder baggrundsartikler, øvelser og videomateriale, samt arbejdsopgaver til timerne og en afleveringsopgave.

På hjemmesiden https://dataanalyseogkraeft.ku.dk findes lærervejledning og forløbsplan.

I 2020, 2021 og 2022 afholdes workshops for undervisere med introduktion til materialet og emnet generelt. Deltagelse i workshop udgør dog ikke en forudsætning for brug af materialet.
Find mere information om workshops på hjemmesiden.

God fornøjelse med materialet!

Projektgruppe

Bioteknologi

Joachim Weischenfeldt, gruppleder, Biotech Research & innovation Centre (BRIC), KU

Aidan Flynn, postdoc, Biotech Research & Innovation Centre (BRIC), KU

Ida Thingstrup, lektor, Roskilde Gymnasium

Estrid Vilma Solyom Høgdall, klinisk professor, Afdeling for Patologi /
Regionernes Bio- og GenomBank (RBGB), Herlev og Gentofte Hospital

Douglas Vinicius Nogueira Perez de Oliveira, Molekylærbiolog, Afdeling for Patologi /

Regionernes Bio- og GenomBank (RBGB), Herlev og Gentofte Hospital

Matematik

Claus Thorn Ekstrøm, professor, Biostatistisk afdeling, Institut for Folkesundhedsvidenskab, KU

Anne Lyngholm Sørensen, Biostatistisk afdeling, Institut for Folkesundhedsvidenskab, KU

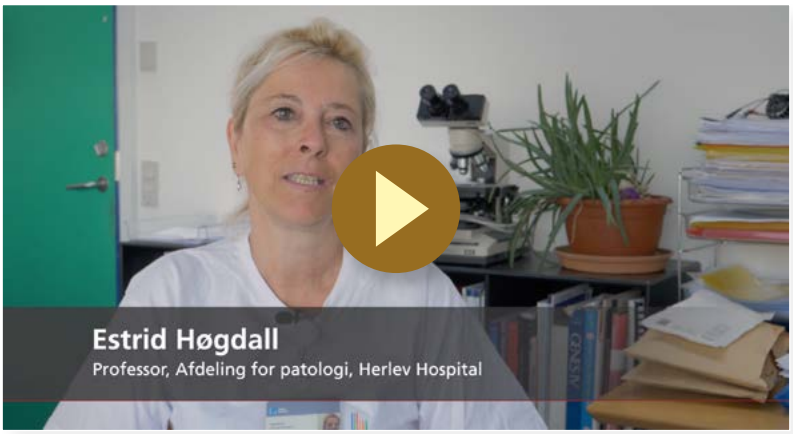
Bjarke Hansen, lektor, Roskilde Gymnasium

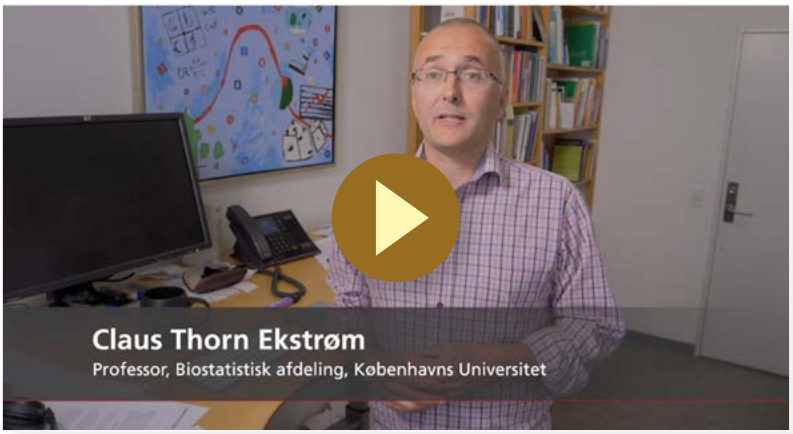
Projektleder:

Anne Rahbek-Damm, communications & outreach officer,
Biotech Research & Innovation Centre (BRIC), KU

Mød eksperterne

I disse videoer kan du høre 3 forskere fortælle, hvordan analyse af genomiske data indgår i deres arbejde.





Kræftbiologi

Aidan Flynn, Ida Thingstrup, Douglas Vinicius Nogueira Perez de Oliveira,
Estrid Vilma Solyom Høgdall, Anne Rahbek-Damm, Joachim Weischenfeldt
| November 2019

Kræftbiologi

Formålet med denne artikel er at introducere en række af de genetiske principper, der ligger bag udviklingen af kræft. Til trods for, hvordan det ofte fremstilles i medierne, er kræft ikke én sygdom men en samlebetegnelse for en lang række af sygdomme, som alle er kendetegnet ved ukontrolleret celledeling. Traditionelt opdeles kræfttyper baseret på den vævs- og celletype, som sygdommen er opstået i, f.eks. melanom fra melanocytter i huden eller brystkræft fra brystvævet. Nye fremskridt inden for den såkaldte genomics-teknologi (f.eks. DNA-sekventering, RNA-sekventering, methyleringsprofilering dvs. epigenetiske undersøgelser i form af hvor DNA er methyleret) viser imidlertid, at der inden for hver kræfttype eksisterer klart definerede undergrupper med hver deres unikke genetiske kendetegn. Omvendt har forskere også observeret sygdomsmønstre på tværs af tilsyneladende ikke-relaterede væv. Af disse grunde giver det almindeligt anvendte udsagn "en kur mod kræft" ikke nogen mening. Hver kræfttype må studeres ud fra sin unikke fysiologi, og for at udvikle effektive behandlingsformer, må vi afdække de mekanismer, der driver udviklingen af forskellige tumortyper.

Udvikling og celleprogrammering

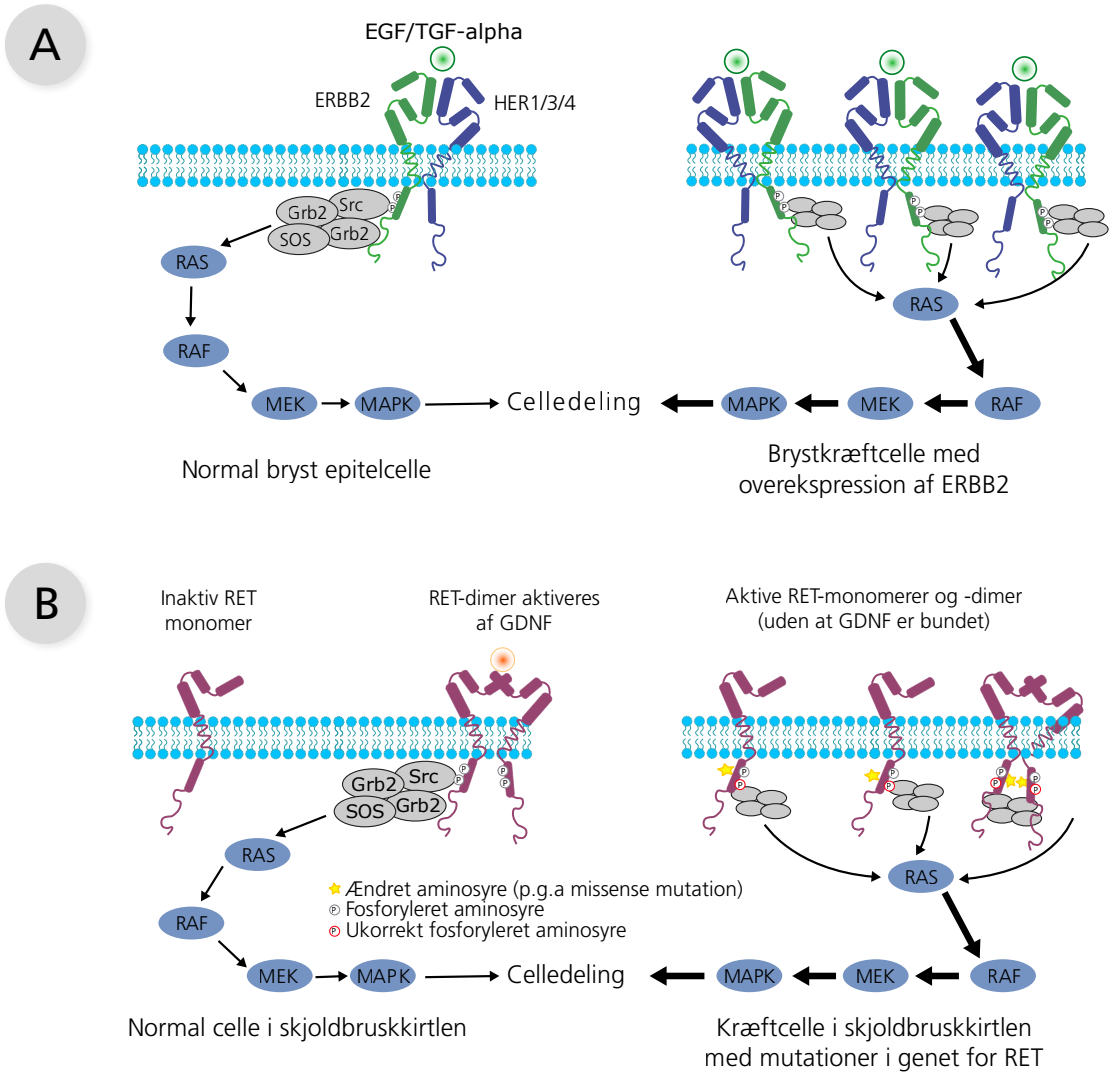
Ethvert menneske begynder livet som en enkelt celle med en unik genetisk kode, der opbevares i vores genom. Udviklingen til et fuldt udviklet menneske indebærer den kontrollerede mangedobling af denne oprindelige celle og dens genom gennem den proces der kaldes *mitose*. Efterhånden som udviklingen skrider frem og organismen bliver stadig mere kompleks, opnår de nye celler en højere og højere grad af specialisering som sætter dem i stand til at udføre specifikke opgaver i vores væv og organer. Den genetiske kode, som fandtes i den oprindelige celle, findes også i alle dattercellerne. Denne genetiske kode indeholder instruktionerne (gener) til opbygning af de maskiner (proteiner), der udfører alle de forskellige opgaver i vores celler. Det er imidlertid ikke alle proteiner der er påkrævet i alle celler. Cellerne specialiseres til specifikke opgaver igennem aktivering af særlige *celleprogrammer.* Et celleprogram styrer, hvilke gener der skal udtrykkes og i hvilket omfang, således at cellen får den rette mængde af de proteintyper, den har brug for, for at kunne udføre de særlige opgaver der kræves af f.eks. en hudcelle, en blodcelle eller et neuron.

Tildelingen af celleprogrammer til forskellige celletyper under udviklingen kræver en meget kompleks kommunikation eller *signalering* mellem cellerne. Når en tumor udvikler sig fra en type af celler (tumor = knude forårsaget af unormal cellevækst), bevarer cellerne fortsat det meste af det celleprogram som er relevant for deres celletype. Dette forklarer delvis, hvorfor vi finder forskellige gener muteret i forskellige tumortyper. For eksempel er det sæt af gener, der forårsager fremkomsten af brystkræft, meget forskellig fra dem der forårsager en hjernetumor, og det betyder, at de forskellige kræftformer ofte kræver helt forskellige behandlinger.

Cellesignalering og vækstkontrol

Hos et voksent menneske udskiftes milliarder af celler hver eneste dag. Det er helt afgørende, at udskiftningen foregår strengt kontrolleret, for hvis forholdet mellem nydannede og døende celler kommer ud af balance resulterer det i sygdomme som kræft. Men hvordan ved en celle, hvornår den skal dele sig, og hvornår den skal dø? Alle celler har et stort antal overfladereceptorer, der bruges til at kommunikere med omgivelserne. Når en receptor kommer i kontakt med det specifikke signalmolekyle, som den "lytter" efter, udløses en proces, så signalet gives videre ind i cellen. Her udløses en kaskade af signalering som kan udløse ændringer i cellen og dermed justere cellens program. Disse signalkaskader kan eksempelvis op- eller nedjustere proliferation (cellevækst og deling), ændre metabolismen (omsætningen af stof eller energi) eller udløse apoptose (selvdestruktion, programmeret celledød). Celler modtager konstant beskeder fra deres omgivelser om alt fra helt normale forhold som eksempelvis en stigning i blodsukker efter et måltid og til ekstreme omstændigheder som eksempelvis en brækket knogle. I det første tilfælde kan muskelceller blive instrueret til at øge deres optagelse af glukose som kan lagres og senere bruges som energi. I det sidste eksempel vil celler, der er ansvarlige for knogledannelse, blive instrueret til at bevæge sig til brudstedet for at dele sig og fremstille nyt knoglevæv, hvilket er en proces, der skal styres præcist, for at undgå misdannelse af den reparerede knogle. En celles liv består altså i at følge det tildelte celleprogram, reagere på eksterne stimuli ved at indgå i faste signaleringsveje og dermed indgå i et stort cellesamfund.

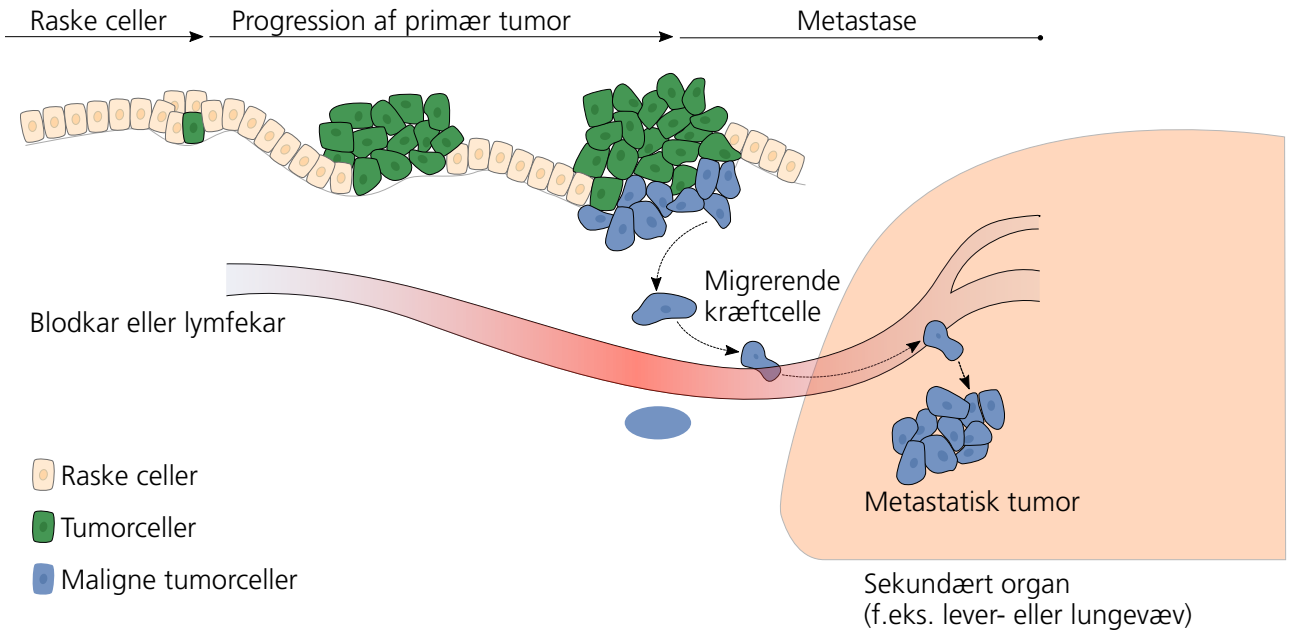
En af de mest almindelige årsager til kræft er overproduktion af særlige vækstsignaler, for eksempel i nogle former for brystkræft, hvor vækstfaktor-receptoren ERBB2 ofte er amplificeret (figur 1A), eller når en receptor bliver permanent aktiveret såsom receptoren RET ved kræft i skjoldbruskkirtlen (figur 1B). Overaktive eller permanent aktiverede receptorer kan instruere celler til at dele sig ukontrolleret, hvilket fører til dannelsen af en tumor.



Figur 1: To virkemåder for onkogen aktivering af receptorer for vækstsignaler.

(A) I normale celler forbindes receptoren ERBB2 med andre receptorer og dette kompleks virker tilsammen som receptor for vækstfaktoren EGF eller vækstfaktoren TGF-α. Når receptoren aktiveres, udløses en signalkaskade som stimulerer celledeling. I nogle brystkræftceller findes der amplifikationer af *ERBB2* genet, som koder for receptoren ERBB2 (som er en tyrosin-protein kinase), hvilket fører til, at der er for meget ERBB2 i cellemembranen. Når der er ekstra af denne receptor, stimuleres der til for stor signalering videre i signalvejene, hvilket i den sidste ende fører til abnorm celledeling. (B) RET-proteinet er inaktivt som monomer, men når det aktiveres af vækstfaktoren GDNF, så samles det til en dimer, som har kinaseaktivitet og som derved aktiverer en signalkaskade, der fører til celledeling. I nogle kræftceller i skjoldbruskkirtlen findes der mutationer i den del af *RET*-genet, der koder for proteinets domæne med kinaseaktivitet. RET-proteinet, der er ændret ved disse mutationer, er i stand til at aktivere den efterfølgende signalkaskade, selvom GDNF ikke binder sig og selvom der ikke dannes en dimer, hvilket i den sidste ende fører til abnorm celledeling.

Tumorer kan være ondartede, hvilket betyder, at tumorcellerne har evnen til at migrere og etablere nye tumorer andre steder i kroppen eller godartede, hvilket betyder, at de ikke har denne evne. Det er kun, når en tumor har ondartet potentiale, at den betragtes som kræft. Se figur 2 (Metastaser).



Figur 2: Metastaser. Kroppens celler og deres funktion, evne til at dele sig og til at bevæge sig rundt er tæt reguleret, og ofte afhængig af det væv de befinder sig i. De fleste celler er ikke mobile og er samtidig afhængige af kommunikation med andre celler i vævet. Tumorceller vil normalt også være afhængige af sådanne signaler, hvilket bl.a. gør at de ikke er i stand til at vandre og bevæge sig ud af vævet. I dette stadie kaldes det en godartet eller benign tumor. Yderligere mutationer kan i visse tilfælde medføre, at tumorcellerne bliver uafhængige af nabo-cellerne, bliver i stand til at nedbryde væv, vandre ud af vævet og over i blod eller lymfebanerne. Langt de fleste tumorceller i dette stadie vil dog dø, men yderligere mutationer kan medføre at tumorcellerne bliver i stand til at vandre ud af blod eller lymfebanerne igen, og ind i andet væv og etablere en ny tumor der. En tumor i dette stadie kaldes malign, og selve processen hvorved tumorceller kan bevæge sig og vandre ud i andet væv kaldes metastase. Begrebet kræft (cancer på engelsk) refererer til tumorer der har erhvervet evnen til at metastasere.

Vedligeholdelse af genomet

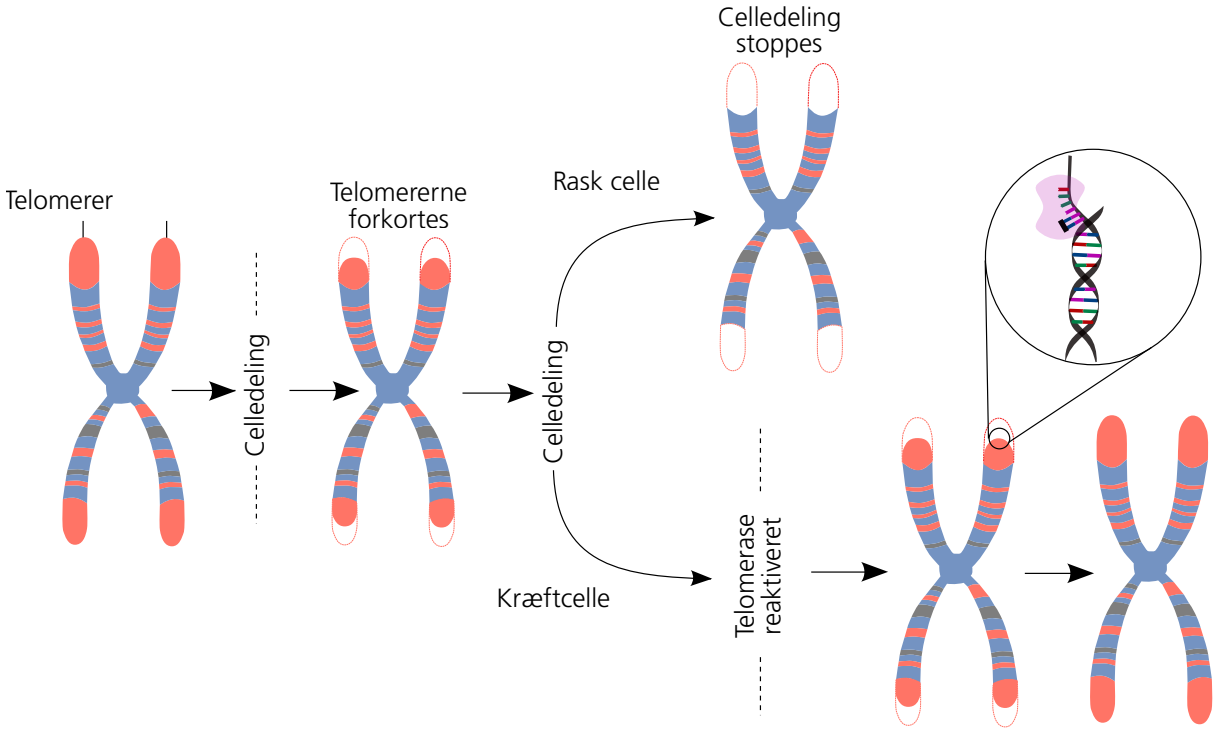
Som tidligere nævnt, begynder alle flercellede organismer livet som en enkelt celle, mens en voksen menneskekrop indeholder milliarder af celler. For at nå dertil skal genomet kopieres nøjagtigt et enormt antal gange. Cellen har et avanceret DNA reparationsmaskineri, som konstant patruljerer genomet for at finde og rette fejl i de tre milliarder nukleotidbasepar. På trods af det enorme vedligeholdelsesarbejde, anslås det, at gennemsnitligt en fejl undslipper reparation per celledeling. En fejl som ikke repareres og som dermed sendes videre til dattercellen, kaldes en *mutation.*

Mutationer kan opstå, når DNA'et ikke kopieres korrekt (såkaldte replikationsfejl) eller de kan forårsages af eksterne faktorer - såkaldte *mutagener*. Mutagener kan f.eks være kemikalier eller stråling, der forårsager ændringer i baserne og deres parring, hvilket gør det svært for cellen at kopiere dem korrekt. I nogle tilfælde kan eksponering endda føre til brud på DNA-strengene. Når vi taler om DNA-skadelige stoffer, tænker vi ofte på ekstreme tilfælde som f.eks. radioaktive udslip eller kemiske våben, men faktisk udsættes vores DNA konstant for potentielt skadelige kemikalier som følge af normal cellulær metabolisme (stofskifte) eller skadelige faktorer i miljøet, såsom UV-lys fra solen. Cellerne har derfor udviklet avancerede systemer som overvåger og reparerer DNA-skader. Reparationerne spænder fra at korrigere en enkelt beskadiget base ved at erstatte den med den rigtige, til at hæfte kromosomer med enkelt-eller dobbelt-strengsbrud sammen. Man mener, at omkring 10.000 baser bliver beskadiget og repareret i hver menneskelig celle hver dag. Og faktisk er kræft ofte forbundet med mutationer i DNA-reparationsenzymer – altså i de proteiner som har til formål at forhindre mutationer!

Når en DNA-skade opstår, binder komplekser af proteiner til det beskadigede DNA. Det bundne kompleks sender besked til andre signalproteiner, som igen bærer meddelelsen til det helt centrale regulatorprotein kendt som Tumor Protein 53 (eller TP53), som også kaldes *genomets vogter*. TP53 aktiverer de nødvendige DNA reparationsveje. I tilfælde hvor skaden på DNA er for omfattende til, at mekanismerne kan rette den, udløser TP53 apoptose. Genet for TP53 er muteret i op mod halvdelen af alle kræftformer og det afspejler, hvor stor betydning genet og proteinet har.

Det er vigtigt at understrege, at langt fra alle mutationer fører til kræft. Faktisk er langt størstedelen af mutationer enten harmløse (såkaldte *passenger mutationer*) eller skadelige for cellen og fører til celledød i stedet for ukontrolleret vækst. Kun når mutationer forekommer i specifikke gener, (hvor de relevante gener til en vis grad afhænger af celletypen som nævnt tidligere) kan de bidrage til tumorvækst. Denne type mutationer betegnes *drivermutationer*. Mutationer i gener som *TP53* kan være enten passenger - eller drivermutationer afhængigt af deres effekt på proteinet.

Da risikoen for, at mutationer integreres i genomet øges hver gang genomet kopieres, har cellerne en naturlig grænse for antallet af gange, de får lov til at dele sig. Hvert kromosom har en beskyttende ende kendt som en telomer, og for hver celledeling bliver telomererne kortere (figur 3). Når cellen registrerer, at telomererne er blevet for korte, holder den op med at dele sig. På den måde beskytter telomerer begge ender af kromosomet og fungerer som et indre "biologisk ur" for cellen. Enzymet telomerase er i stand til at forlænge telomererne, men det er deaktiveret i de fleste modne celler. Ved mange kræftformer er telomerasegenet blevet genaktiveret. Det medfører, at telomerasen forhindrer, at kromosomernes telomerer afkortes under den forøgede celledeling.

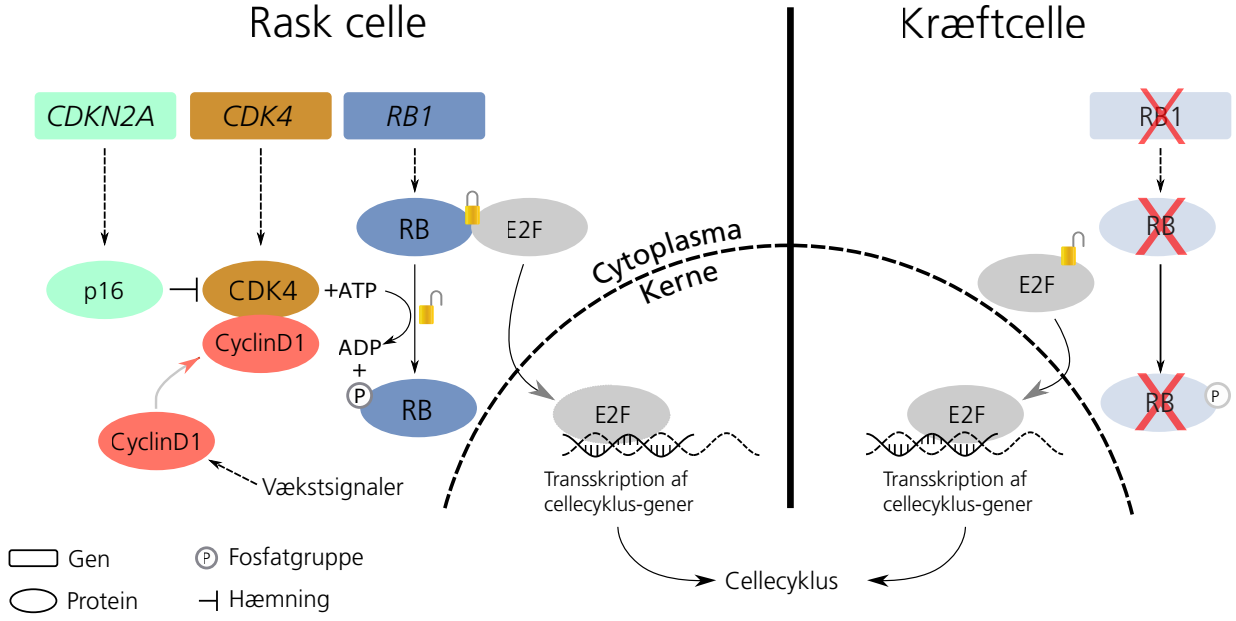


Figur 3: Telomerase. Kromosomer bliver replikeret under cellecyklus. Stort set hele genomet bliver nøjagtigt replikeret undtagen enderne (telomererne), som bliver afkortet en lille smule hver gang en celle deler sig. Når telomererne når en kritisk kort længde, typisk efter ca 40-50 celle-delinger, går cellerne i stå (senescence) eller undergår programmeret celledød (apoptose). Telomererne har yderligere en særlig protein-kappe, som sørger for at kromosomer ikke fusionerer, da cellen normalt vil forsøge at reparere dobbelt-strenget DNA-ender. Uden denne protein-kappe, vil der være stor risiko for at kromosomerne bliver ligeret sammen. Celler har et særligt enzym kompleks, telomerasen, som kan replikere telomererne, dvs forlænge dem således, at de ikke forkortes. Telomerasen kan aktiveres, når der er behov for det, som for eksempel i embryonale stamceller, der skal dele sig mange gange. Tumorceller er karakteristiske ved at dele sig uhæmmet, og aktivering af telomerase er derfor et typisk træk ved tumorceller.

Kræft

Kræft er resultatet af en manglende regulering og kontrol af cellerne. Gener, som i muteret tilstand forårsager en stigning i vækst kaldes onkogener, mens gener som sikrer sikkerhedsmekanismer i cellen kaldes tumorsuppressorgener (TSG). Samlet set kaldes onkogener og TSG'er for *drivergener*. En vigtig pointe er, at mens drivermutationer altid forekommer i drivergener, er det ikke alle mutationer som forekommer i drivergener som er drivermutationer.

To tumorsuppressorgener, hvis proteiner ofte mister deres funktion i forbindelse med kræft er *TP53* og *RB1*. Begge proteiner spiller en vigtig rolle for reguleringen af celledeling. Som tidligere beskrevet er TP53 ansvarlig for at overvåge DNA'et og forhindre cellen i at undergå celledeling, før eventuelle skader er udbedret. RB1 proteinet virker tilsvarende ved at blokere for celledeling. Først når proteinet modtager et tilstrækkeligt stærkt vækstfremmende signal via celleoverfladereceptorer, løsner det blokeringen og tillader cellen at dele sig (se figur 4).



Figur 4. Retinoblastom (RB)-proteinet er hovedregulator af cellecyklus.

Retinoblastom (RB)-proteinet er en hovedregulator af cellecyklus i den normale celle: Venstre side af figuren: Transskriptionsfaktoren E2F er en aktiverende transskriptionsfaktor for gener, hvis proteinprodukter stimulerer cellecyklus. Når proteinet RB ikke er fosforyleret, hindrer det at cellen går videre i sin cellecyklus. Det sker ved, at RB fastholder E2F og derved hindrer E2F i at fremme transskription af gener, der stimulerer cellecyklus. Når cellen modtager et vækstfremmende signal, vil CyclinD1 binde sig til CDK4 og dette kompleks kan fosforylere RB-proteinet. Den fosforylerede udgave af RB kan ikke binde og hæmme E2F, som nu er fri, og derved kan E2F fremme transskriptionen af cellecyklus-generne. Flere proteiner påvirker denne regulering, idet proteinet p16 (udtrykt af genet *CDKN2A*), virker som en bremse på celledelingen i den normale celle. Det sker ved, at p16 bindes til CDK4 og derved hæmmer CDK4's binding med CyclinD1. Når CyclinD1 hæmmes i at blive bundet til CDK4, hæmmes også fosforyleringen af RB, og dermed hæmmes celledelingen. I denne figur ses altså et eksempel på den komplicerede balance mellem fremmende og hæmmende signaler til cellecyklus. Højre side af figuren: Både generne *RB1* og *CDKN2A* er ofte muteret eller helt deleteret (slettet) i kræftceller. figuren viser, hvordan deletion af *RB1* fører til konstant aktivering af cellecyklus.

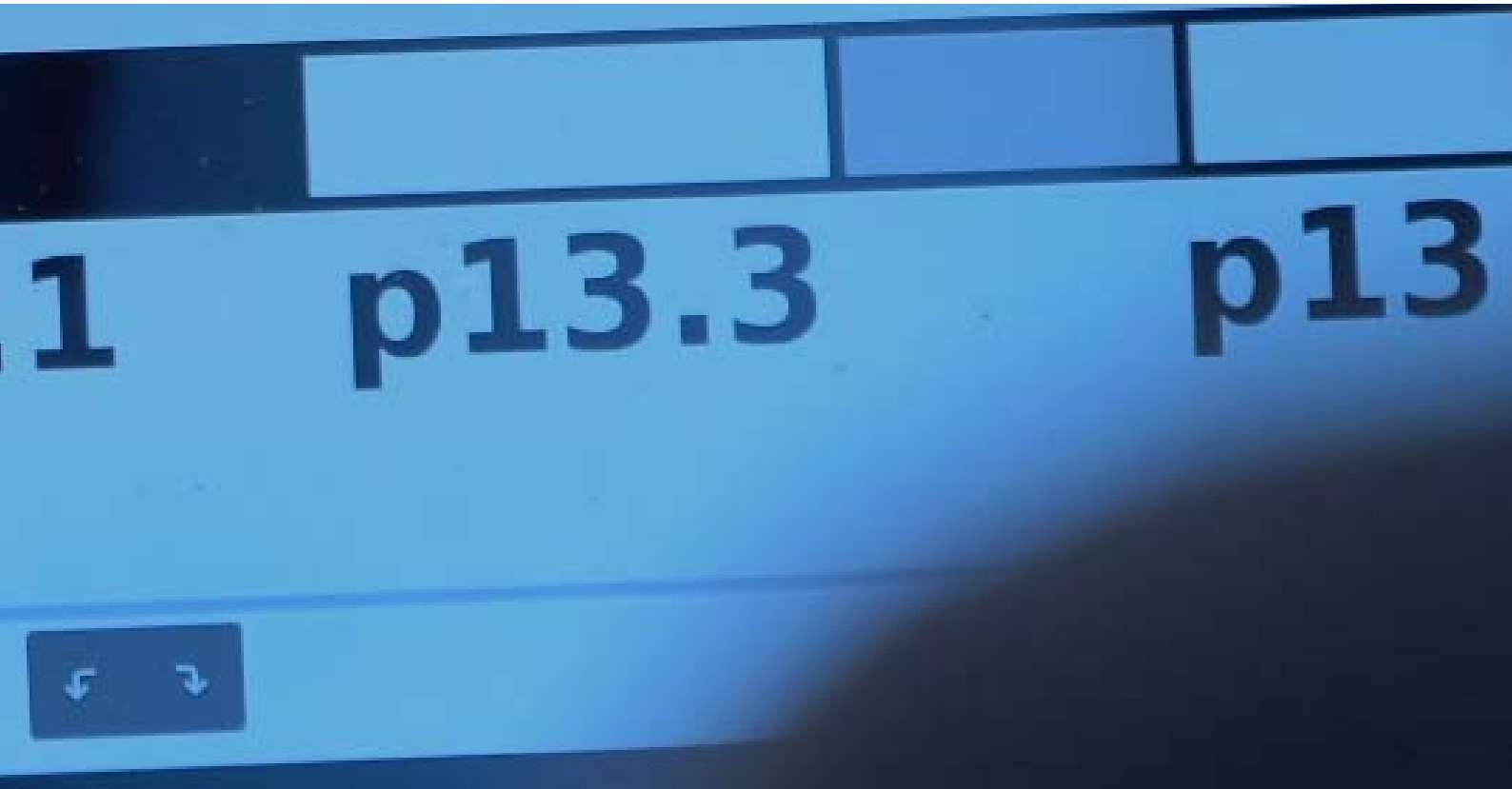
Hvis et af de to proteiner mister deres funktion, mister cellen et vigtigt sikkerhedscheckpoint, som forhindrer ukontrolleret vækst. Når "overvågnings proteinet" TP53 sættes ud af spil, er der ikke noget til at forhindre mutationer i at opstå i andre vigtige gener, og cellen drives dermed frem mod kræfttilstand. I kræftceller, hvor den normale kontrol over celledelingen er tabt, kan man ofte konstatere store skader netop på TSG'er i form af deletion. Det betyder, at proteinerne, som generne koder for, enten ikke virker korrekt, eller slet ikke bliver dannet, og dette har bidraget til udviklingen af kræft. Mennesker har to kopier af hvert gen, men én kopi er ofte tilstrækkelig til at udføre genets funktion. I kræftceller er det derfor almindeligt at se at begge kopier af TSG'er er ramt af mutation, deletion eller en kombination af begge. Dette forklarer også, hvorfor mennesker, der er født med en mutation i et TSG (for eksempel *BRCA1*-genet, som er muteret ved nogle former for bryst- og æggestokkræft) overordnet set er sunde, men mere modtagelige for kræft, fordi deres celler ikke har en "ekstra" kopi af genet, som kan stå for funktionen alene, hvis den anden kopi også muterer.

I modsætning til tumorsuppressorgenerne, er onkogenernes rolle i kræft (via de proteiner, som de koder for) at drive cellecyklus fremad og fremme celledeling i for høj grad. Onkogener er muterede varianter af normale gener, som ofte indgår i cellesignaleringsveje og modtager eller spreder vækstfremmende signaler i cellen. De normale varianter af generne kaldes proto-onkogener og de muterede udgaver som man ofte finder i tumorer er altså onkogener. Et eksempel på en signalvej, der påvirkes ved kræft er den

såkaldte Mitogen-aktiverede-protein-kinase (MAPK er navnet på et centralt protein, som ses på figur 1) signalvej. Nogle proto-onkogener er kinaser som er en type enzym, der fosforylerer (dvs. tilføjer en fosfatgruppe til) specifikke aminosyrer på et andet protein. Fosforyleringen omdanner ofte proteinet fra en inaktiv til en aktiv form. Mange celleoverfladereceptorer er kinaser, som er inaktive, indtil det passende signalmolekyle binder sig. Den aktiverede receptor fosforylerer og aktiverer derefter effektorproteiner, som igen fosforylerer andre effektorproteiner. Denne kaskade af signalering udløser i sidste ende den ændring i cellen (f.eks. ændret ekspression af et specifikt gen), som er nødvendig for at udføre den opgave, som det modtagne signalmolekyle var en instruks til. Denne opgave er for proto-onkogenernes vedkommende at stimulere normal vækst og celledeling.

Når der forekommer mutationer i proto-onkogener, kan resultatet blive et onkogen, der f.eks. koder for en ændret kinase, som er permanent aktiv i modsætning til den normale udgave, som kun er aktiv, når den modtager instruks om det. Dermed overaktiveres signalvejen og det kan føre til kræft. Man kan bruge sammenligningen med en speeder, der sidder fast, så bilen hele tiden kører hurtigt, og i lighed med dette bliver cellens delingsaktivitet også overstimuleret. I den enkelte tumor er det usædvanligt at se mere end ét muteret onkogen per signaleringskaskade. Det er nok at ét gen ændres til overstimulering, da det er ligegyldigt for slutresultatet, nemlig ukontrolleret celledeling, om signalet bliver tvangsaktiveret i starten, midten eller slutningen af signaliseringskaskaden. Så det er nok med én mutation i en signaleringskaskade for at celledelingen ender med at blive overstimuleret via denne kaskade, og cellen kan omdannes til en kræftcelle uden yderligere tilfældige mutationer i andre gener i den samme kaskade.

Mutationer er ikke den eneste måde, hvorpå onkogener kan blive overaktiveret. Såkaldt gen-amplifikation, hvor antallet af kopier af et onkogen er forøget fra den normale diploide tilstand (to kopier, ét på hvert af de to homologe kromosomer) til mange flere kopier på et eller begge kromosomer, kan ligeledes føre til forøgede niveauer af proteinet og dermed øget signalaktivering. Dette ses oftest med gener der koder for receptorer for vækstfaktorer såsom *EGFR* ved hjernekræft eller *ERBB2* ved brystkræft (se figur 1A).



Kræft i kontekst

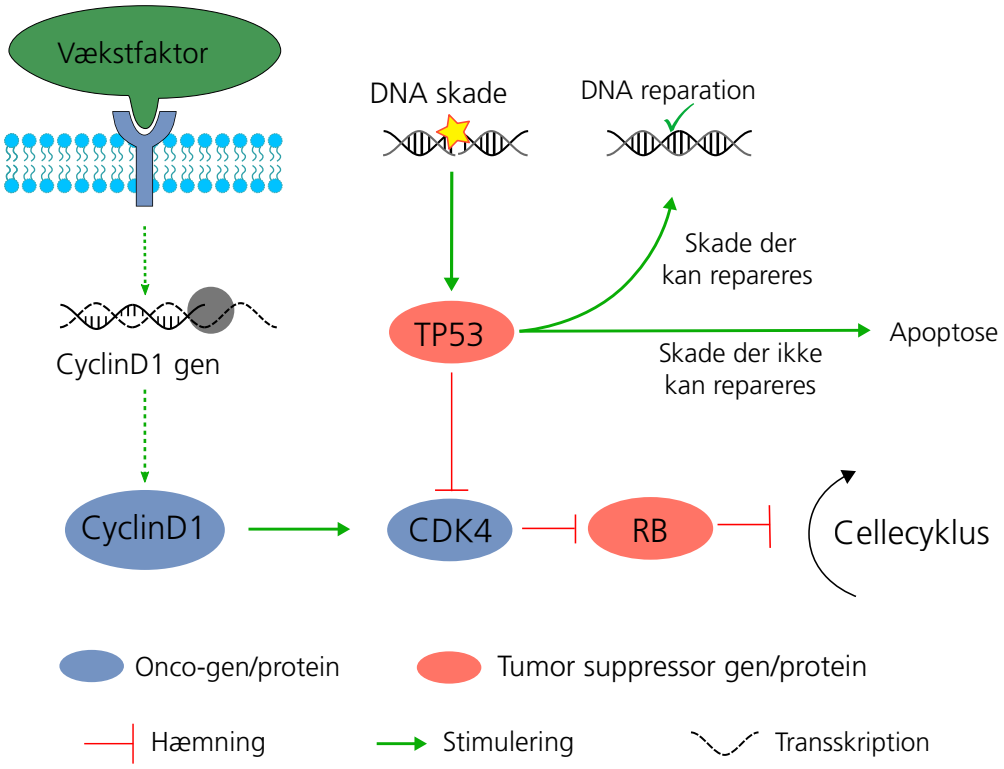
I 2000 definerede to fremtrædende kræftforskere seks afgørende kendetegn ved kræft; 1) Konstant aktiveret vækstsignalering, 2) evnen til at undgå apoptose, 3) ubegrænset celledelingsevne, 4) evnen til at undgå vækstthæmmere 5) evnen til at fremkalde angiogenese (dannelse af nye blodkar) og 6) evnen til at aktivere invasion af andre væv og lave metastaser.

Denne artikel beskæftiger sig hovedsageligt med de tre første kendetegn, som vedrører de processer, der finder sted i den enkelte kræftcelle. Kræftceller eksisterer imidlertid ikke isoleret. De sidste tre kendetegn omhandler kræftcellers samspil med resten af kroppen og betydningen af dette stiger i takt med at en tumor vokser. For at skaffe den store mængde energi og næringsstoffer der kræves af en voksende tumor, rekrutterer kræftcellerne normale endotelceller (det lag af celler som dækker blodkarrenes inderside) for at opbygge nye blodkar ind i tumoren. En succesfuld kræftcelle skal også undvige immunsystemet, som konstant forsøger at ødelægge unormale celler. En af de måder hvorpå kræftceller opnår dette er ved at opregulere dannelsen af immunundertrykkende markører, som forhindrer immunsystemet i at genkende og angribe kræftcellerne.

Et par af de mest vellykkede immunterapeutiske behandlinger som er udviklet i løbet af de senere år er baseret på blokering af disse immunundertrykkende markører, så patientens immunforsvar igen kan virke mod de unormale celler. På samme måde har lægemidler, som blokerer angiogenese – altså dannelse af blodkar, været succesfulde i behandlingen af visse typer kræft. Dette er blot to eksempler på det komplekse samspil mellem en tumor og dens omgivende miljø, og på hvordan disse kan blive et mål for behandling.

Konklusion

Kræft skyldes – helt kort sagt – fejlregulering af cellevækst. Der er imidlertid mange sikkerhedsforanstaltninger og miljømæssige udfordringer, som en celle skal overvinde, før den kan blive til en kræftcelle. De mekanismer som fører til sygdom er primært ændringer i cellens genetiske kode, hvilket resulterer i afvigende proteiner, der enten kaprer eller ødelægger de normale signalveje i en celle. Det er ofte de centrale signalveje som er involveret i cellevækst, celleoverlevelse og kontrolpunkter i cellecyklus, der berøres (se figur 4 og 5). Ved at afdække hvordan disse signalveje aktiveres, har forskere været i stand til at udvikle en række effektive lægemidler, som kan ramme specifikke overaktive signalveje og afbryde signaleringen. Yderligere forståelse af de genetiske mekanismer, der er ansvarlige for kræftdannelse, er en integreret del af udviklingen af nye diagnostiske og terapeutiske strategier.



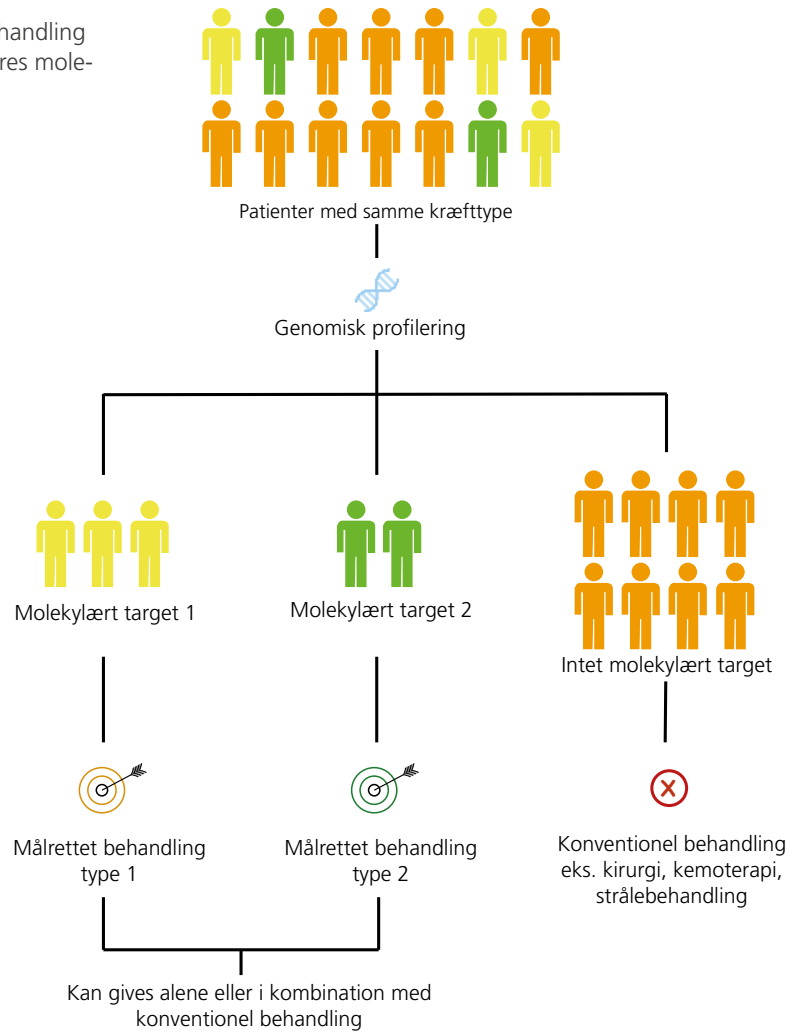
Figur 5: Samspil mellem signalveje står for en kompleks kontrol af cellernes funktion. DNA skader kan opstå som følge af mutagener (f.eks. UV stråling) eller når DNA ikke bliver korrekt replikeret. Celler har et sofistikeret sæt af enzymer der kan reparere de fleste DNA-skader, men mutationer opstår, når skaden ikke bliver korrekt repareret. I tilfælde af en DNA skade, vil et sæt af DNA-skade respons proteiner binde til DNAet, og disse vil derefter aktivere en stribe proteiner, heriblandt TP53 (som kodes for af tumorsuppressorgenet *TP53*). Drejer det sig om få skader på DNAet, vil TP53 aktivere et sæt af enzymer der kan reparere og udbedre DNA skaden. Det er essentielt for cellen, at DNAet ikke bliver replikeret, før skaden er udbedret, og TP53 vil derfor samtidig hæmme cellecyklus ved bl.a. at hæmme CyclinD1/CDK4. Dette kompleks, der ellers normalt fosforylerer RB og derved normalt stimulerer cellecyklus (se figur 4) bliver altså forhindret i at fosforylere RB, som dermed hæmmer cellecyklus. Denne hæmning sker på trods af, at en vækstfaktor har bundet sig til receptoren og har igangsat en signalkaskade til fremme af cellecyklus. I tilfælde af ekstrem DNA skade (som f.eks. ved udsættelse for høje mængder UV-stråling), vil TP53 aktivere programmeret celledød (apoptose). På grund af den helt centrale rolle for *TP53* i at kontrollere celledeling og reparation, er det ikke overraskende at *TP53*er det hyppigst muterede gen i kræft (ca. 50% af alle kræftformer har mutationer i *TP53* genet).

Præcisionsmedicin i kræftbehandling

Aidan Flynn, Ida Thingstrup, Douglas Vinicius Nogueira Perez de Oliveira,
Estrid Vilma Solyom Høgdall, Anne Rahbek-Damm, Joachim Weischenfeldt
| November 2019

Begrebet præcisionsmedicin bruges generelt om behandling som er tilpasset den enkelte patient på baggrund af hans/hendes molekylære profil. Selvom begrebet er moderne, er ideen bag præcisionsbehandling langt fra ny – blandt de tidligste kendte eksempler hører den første succesfulde blodtransfusion baseret på blodtype som fandt sted i 1818. I dag tilbydes målrettet behandling inden for en bred vifte af sygdomme, takket være teknologiske fremskridt og øget forståelse af de molekylære aspekter ved forskellige sygdomme.

De fleste har nok hørt begreberne Præcisionsmedicin, Personlig medicin eller Individualiseret medicin. De forskellige udtryk bruges alle til at beskrive samme koncept. Udtrykkene "personlig" og "individualiseret" kan give det indtryk, at der skræddersyes en helt unik behandling til hver enkelt patient, hvorimod begrebet Præcisionsmedicin kommer tættere på at beskrive det praktiske mål: nemlig at matche



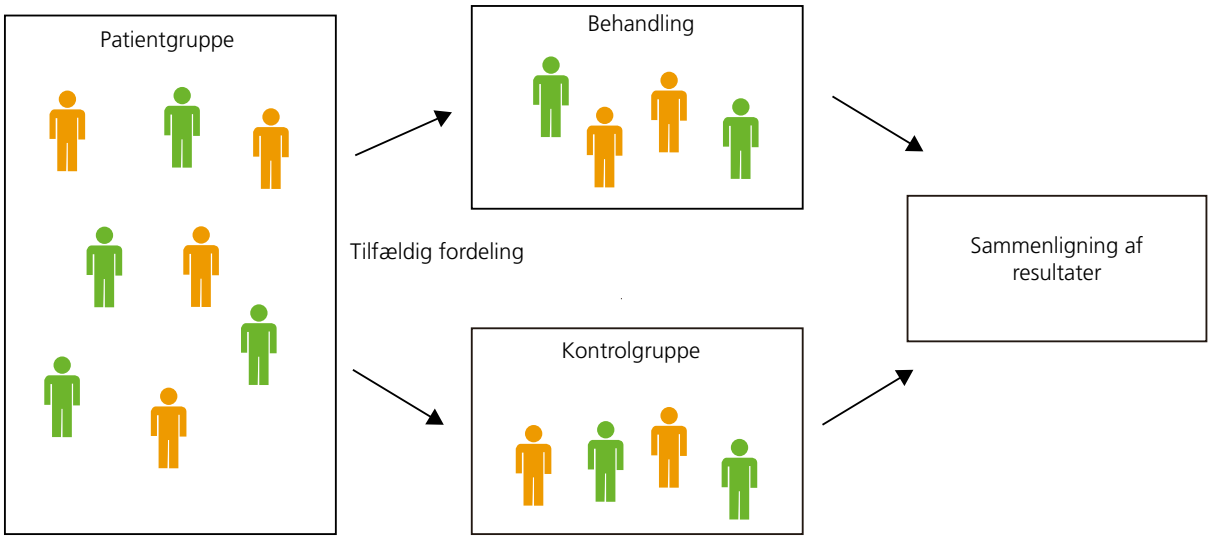
Figur 1. Skræddersyet behandling af patienter baseret på deres molekylære profil

patienter med de mest effektive behandlinger ved at gruppere patienterne baseret på deres genetik, livsstil og miljømæssige faktorer (figur 1).

I dag eksperimenteres der med præcisionsmedicinske tiltag inden for en lang række sundhedsområder – lige fra forebyggelse af diabetes til dosering af lægemidler, men ingen steder er udviklingen og implementeringen af præcisionsmedicin gået så stærkt som inden for kræftbehandling. Det skyldes ikke mindst, at læger og forskere inden for dette felt har haft adgang til enorme mængder genomisk information om de molekylære mekanismer, som ligger til grund for udviklingen af kræft. I takt med den øgede forståelse af sygdomsmekanismerne, har videnskaben været i stand til at udvikle målrettede behandlingsformer, der dræber eller hæmmer kræftcellerne med færre bivirkninger.

Præcisionsmedicin: hvorfor, hvordan, hvornår

På mange hospitaler benytter læger i dag patienters genetiske information til at kortlægge familiens sundhedshistorie og sygdomsrisici. Denne tilgang står i kontrast til den traditionelle "one-size-fits-all" -tilgang, hvor alle patienter med samme sygdom modtager den samme type behandling. Konventionelle kræftbehandlinger er baseret på såkaldte randomiserede kontrollerede forsøg, hvor store patientgrupper inddeles på basis af nogle få karakteristika, for at teste effektiviteten af en specifik behandling (eksempelvis om en specifik kemoterapi giver længere overlevelse for en bestemt type kræft sammenlignet med patienter, der ikke har modtaget kemoterapi, figur 2). I præcisionsmedicin for kræftbehandling er disse store, kontrollerede undersøgelser ofte ikke mulige, da behandlingen er baseret på flere specifikke molekylære ændringer hos den enkelte patient.

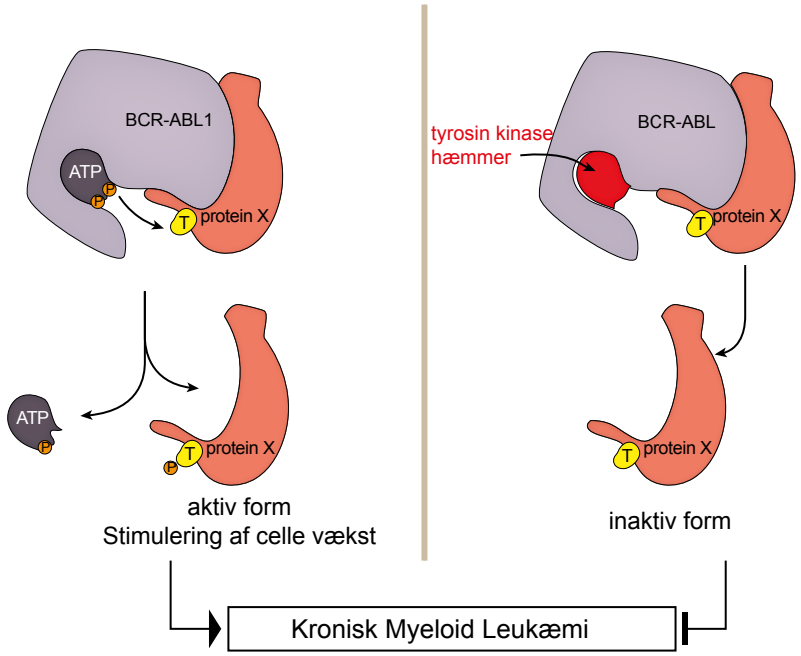


Figur 2. Randomiseret, kontrolleret forsøg som led i "klassisk" lægemiddel afprøvning

I baggrundsartiklen om kræftbiologi blev det beskrevet, hvordan ukontrolleret cellevækst er et nøgleelement i kræft, samt betydningen af såkaldte tumorsuppressorgener og onkogener. Formålet med kræftpræcisionsmedicin er at bruge molekylær information til at identificere og ramme præcis de ændrede signalveje, som er ansvarlige for den ukontrollerede cellevækst. Tumorsuppressorgener bliver typisk tabt eller inaktiveret i forbindelse med kræft, og det gør dem vanskelige at benytte som mål for lægemidler, da de fleste lægemidlers effekt består i at blokere eller inaktivere målproteinet. Onkogener bliver derimod som regel opreguleret eller muteret således, at de får andre egenskaber end den normale version af proteinet. Det gør dem til langt mere oplagte mål for lægemidler, da det er lettere at slukke for et overaktivt enzym (onkogen) end at aktivere et slettet (tumorsuppressor) gen.

En succeshistorie

Behandling af kronisk myeloid leukæmi (CML) udgør et eksempel på en vellykket behandling, der er baseret på molekylær information. Før i tiden behandlede man CML med knoglemarvstransplantation, hvilket er forbundet med alvorlige bivirkninger. Dette ændrede sig, da forskere identificerede en kromosom- translokation (også kendt som Philadelphia-kromosomet) mellem *BCR* genet og tyrosin-kinase genet *ABL1*, som er til stede i ca. 90% af alle CML-tilfælde. Translokationen medfører en permanent aktiv form af ABL1 proteinet. BCR-ABL1-fusionen banede vejen for en af de første og mest succesrige målrettede kræftbehandlinger: nemlig et lægemiddel som retter sig specifikt mod fusions tyrosin-kinasen (figur 3). Lægemidlet hæmmer kinaseaktiviteten af ABL1, og udgør i dag en del af standard-behandlingen for CML-patienter såvel som andre kræftformer med aktivering af lignende tyrosin-kinaser.



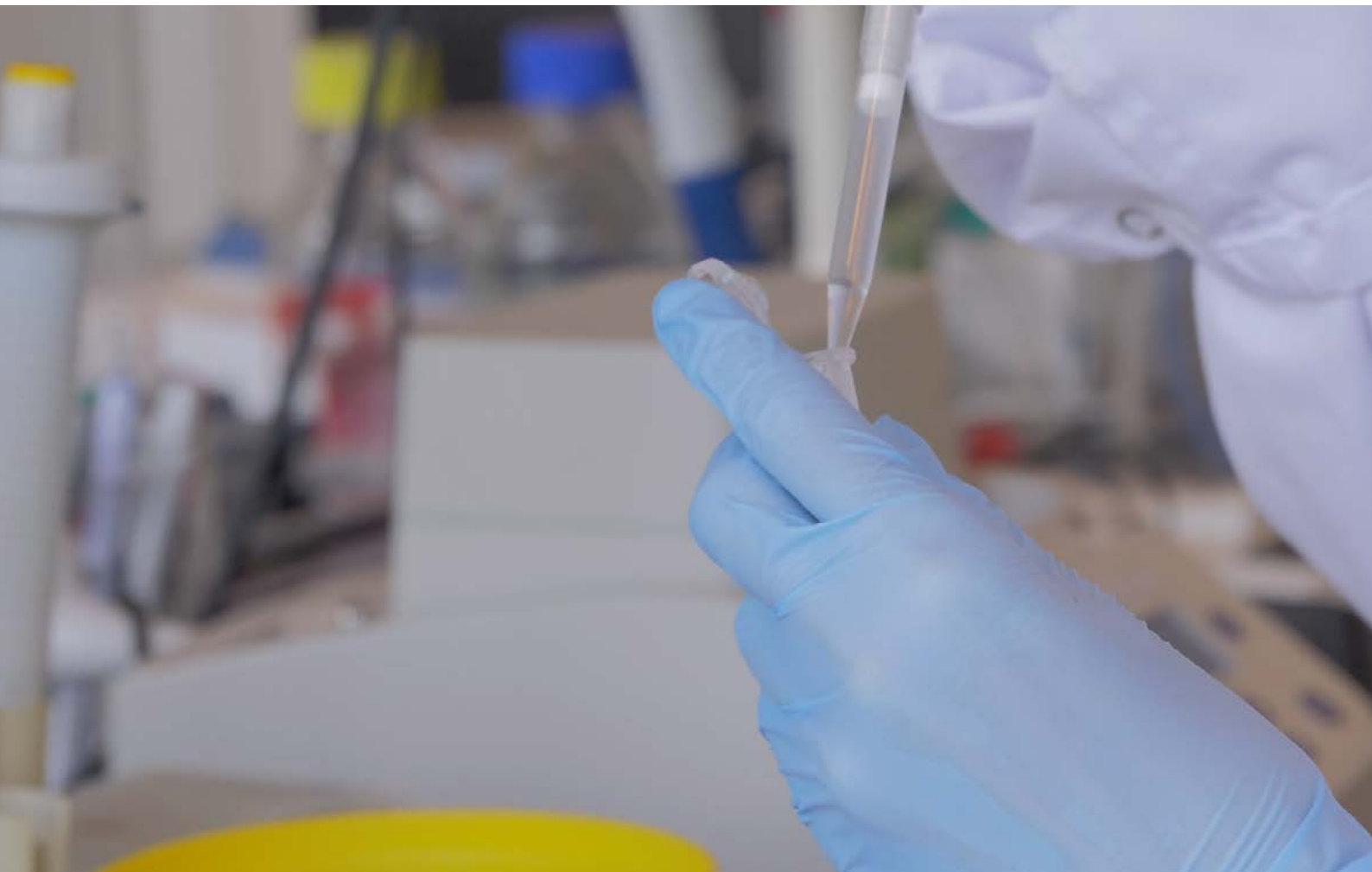
Figur 3. Fusions-proteinet BCR-ABL1 er en hyper-aktiv tyrosin kinase, der phosphorylerer Tyrosin amino syrer på mål-proteiner ("protein X"), der derved bliver aktiveret og kan stimulere f.eks. Ras signallerings kaskade. En tyrosin-kinase hæmmer (f.eks. imatinib) binder til det aktive domæne på ABL1 og medfører at ABL1 ikke længere kan phosphorylere mål-proteiner.

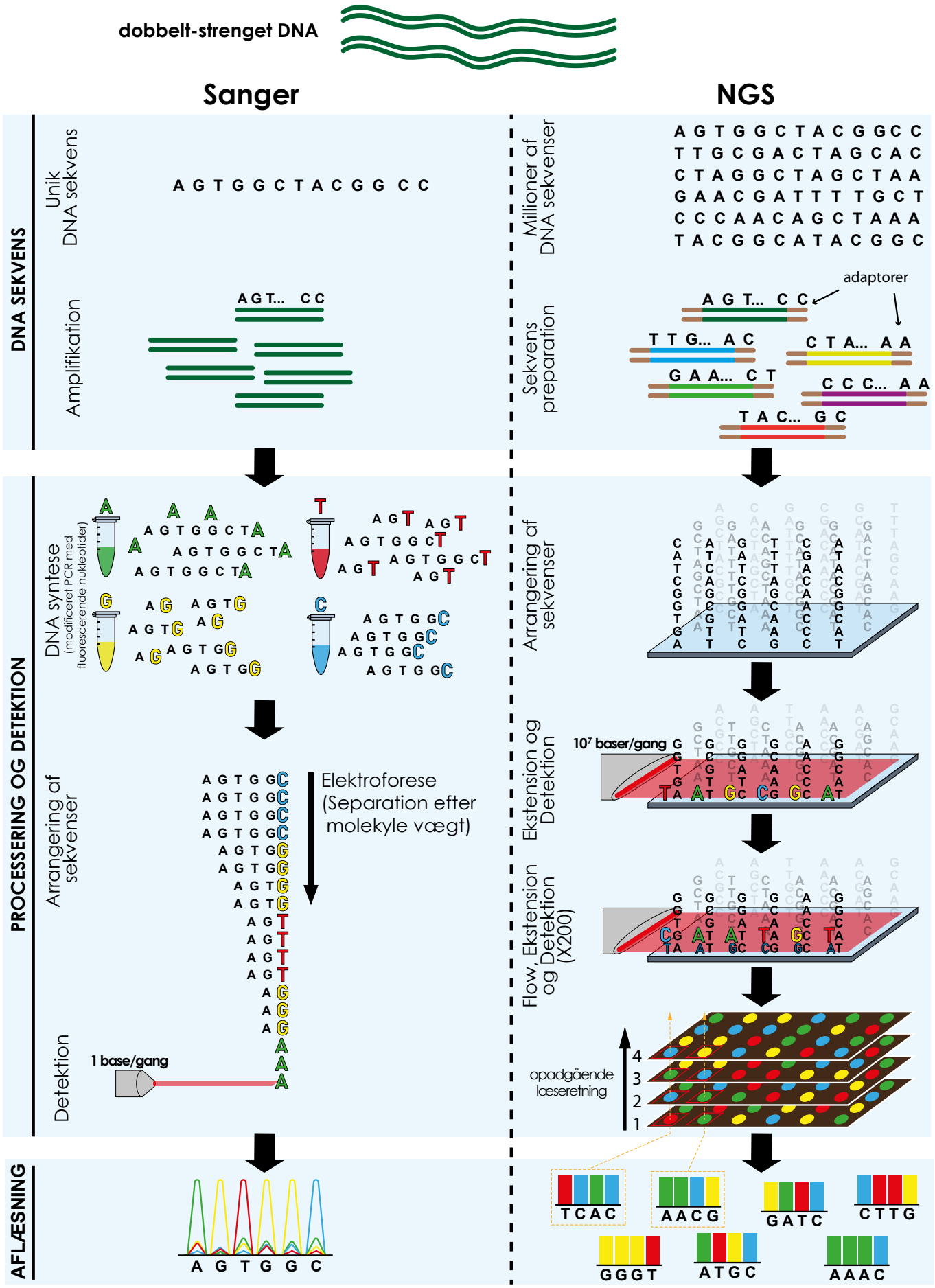
På trods af, at dette lægemiddel har revolutioneret CML behandlingen, kan behandlingsresistens forekomme - ofte på grund af mutationer i ABL1's kinasedomæne, der er nødvendige for bindingen af lægemidlet (figur 3). Kampen for at overvinde resistens har ført til udviklingen af både anden, tredje og fjerde generation af tyrosin-kinasehæmmende lægemidler, og dermed illustrerer sagen også den "evighedskrig" der udspiller sig mellem kræftsygdomme og udviklere af præcisionsmedicin.

Teknologiske fremskridt inden for sekventering

De store fremskridt inden for præcisionsmedicin skyldes ikke mindst opfindelsen af nye teknologier til DNA sekventering - såkaldte high throughput technology. Samtidig er disse teknologier blevet så meget billigere at anvende, at de nu kan bruges rutinemæssigt i det offentlige sundhedsvæsen. Da det første menneskelige genom blev sekventeret i 2001 var prisen næsten 20 milliarder kroner, mens det i dag er muligt at sekventere et helt menneskeligt genom på et par dage for omkring 6.000 kr. Lægemiddelindustrien var hurtige til se og udnytte potentialet i sekventering, og anvender i dag sekventeringsteknologier på forskellige måder i forbindelse med udviklingen af præcisionsmedicin.

Den nye type sekventeringsteknologi som for alvor har revolutioneret DNA sekventering kaldes "Next Generation Sequencing" (NGS). Metoden er en videreudvikling af den klassiske Sanger-sekventering med den forskel, at NGS ikke blot aflæser ét men mange millioner små DNA-fragmenter på samme tid (figur 4). NGS kan udføres på stort set samme tid som Sanger sekventering og til en brøkdel af prisen. NGS gør det også muligt at foretage mere detaljerede og sensitive undersøgelser af de forskellige typer af DNA-ændringer, som eks. forskellige typer af mutationer, deletioner, amplifikationer og andre typer af genomiske rearrangeringer.




dobbelt-strenget DNA
Sanger
NGS
DNA SEKVENS
Unik DNA sekvens
A G T G G C T A C G G C C
Amplifikation
A G T... C C
Millioner af DNA sekvenser
A G T G G C T A C G G C C
T T G C G A C T A G C A C
C T A G G C T A G C T A A
G A A C G A T T T T G C T
C C C A A C A G C T A A A
T A C G G C A T A C G G C
Sekvens preparation
adaptorer
A G T... C C
T T G... A C
C T A... A A
G A A... C T
C C C... A A
T A C... G C
PROCESSERING OG DETEKTION
DNA syntese
(modificeret PCR med fluorescerende nukleotider)
Arrangering af sekvenser
Elektroforese
(Separation efter molekyle vægt)
Detektion
1 base/gang
Arrangering af sekvenser
Ekstension og Detektion
10⁷ baser/gang
Flow, Ekstension og Detektion (X200)
opadgående læseretning
4
3
2
1
AFLÆSNING
A G T G G C
TCAC
AACG
GATC
CTTG
GGGT
ATGC
AAAC

Figur 4.

Sekventeringsmetoder: Sanger vs NGS. Illustrationen viser de vigtigste forskelle mellem Sanger-sekventering (til venstre) og et eksempel på Next Generation Sequencing (NGS) (til højre, mere konkret vises pyrosekventering, som er en af flere mulige metoder inden for overbegrebet NGS). Ved Sanger-sekventering kan man kun sekventere ét DNA-fragment ad gangen, mens NGS kan sekventere millionvis af DNA-fragmenter på samme tid. NGS benytter en "shotgun sequencing" tilgang, hvor der sekventeres mange små stykker DNA, som i forvejen er blevet klippet i små DNA fragmenter med enzymer. Små stykker DNA kaldet 'adaptorer' sættes på alle DNA-fragmenterne, således at de kan anbringes på en lille glasplade. På glaspladen er på forhånd anbragt korte DNA-stykker, som er komplementære til adaptorerne, og som er kemisk bundet fast til glaspladen. Enkelt-strengede DNA fragmenter bliver på den måde hæftet på glaspladen vha base-parring mellem adaptorerne på DNA fragmenterne og adaptorerne på glaspladen. Derefter sker der i princippet nogenlunde det samme ved begge metoder: Der tilsættes DNA-polymerase, primere og de nødvendige nukleotider (A, T, C og G), som vha polymerase-reaktion, danner den komplementære streng til de enkeltstrengede DNA-stykker, som er hæftet på glaspladen. Der tilsættes en blanding af flourescensmærkede nukleotider af hver slags (A, T, C og G, med en særlig farve for hver) som desuden er kemisk modificeret, så DNA-syntesen stopper, når et af disse ændrede nukleotider indsættes. Forskellen er, at der ved Sangersekventering kun sekventeres ét DNA-fragment per reaktion, mens der med NGS kan sekventeres rigtig mange per glasplade, fordi hver enkelt sekvens sidder i små klynger på glaspladen, og der er plads til mange klynger på glaspladen.Sekvensen bestemmes ud fra rækkefølgen af hver fluorescerende nukleotid:For Sanger-sekventeringen køres en gel-elektroforese med produkterne fra syntesen, hvor de mindste stykker bevæger sig hurtigst igennem gelen og de største bevæger sig langsomst. Man lader prøverne løbe ud af gelen i bunden, hvor der sidder en meget følsom laserdetektor, som hele tiden registrerer i hvilken rækkefølge DNA-stykker med de forskellige fluourescerende farver kommer igennem gelen.

Ved NGS sker detektionen samtidig for alle DNA fragmenter én nukleotid ad gangen. Efter hver fluorescerende nukleotid er tilsat og bundet, skylles reagenserne ud maskinelt, og farven detekteres på glaspladen i hver klynge på et billede med høj opløsning. Nukleotiderne C, T, A og G kan f.eks. være hæftet på blåt, rødt, grønt og gult fluoreserende farve. Alle de DNA fragmenter hvor f.eks. et "C" er indsat vil således udsende blåt lys. Dernæst fjernes den kemiske blokering på hver nukleotid, således at en ny nukleotid kan indsættes. Endelig tilsættes igen DNA-polymerase og nukleotider. Dette fortsætter mellem 50 og 300 gange, afhængig af hvilket system, der er tale om. Sekvensen for hvert oprindeligt DNA-fragment på glaspladen kan nu aflæses ved at sammenstykke informationen fra alle billederne, som blev taget i løbet af processen. Dette vil give mange millioner af DNA-sekvenser på mellem 50 og 300 nukleotider, og det næste trin er bioinformatisk at finde, hvordan disse fragmenter hører sammen.

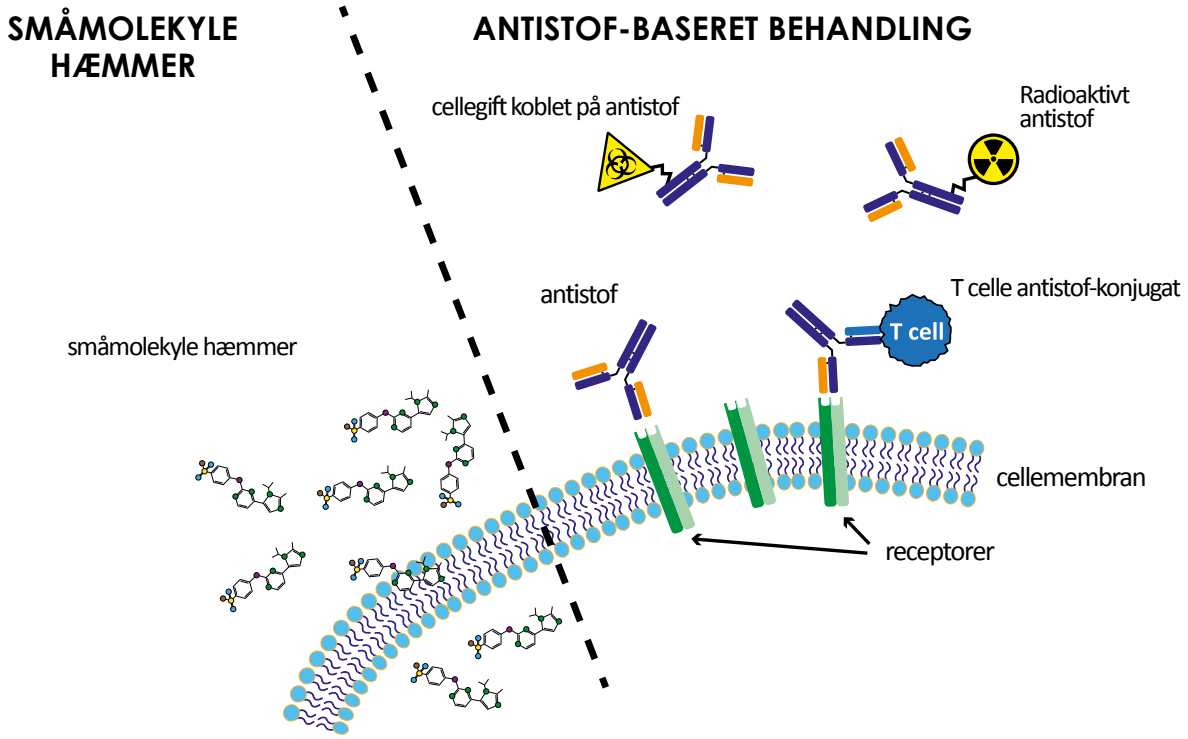
Muligheder og udfordringer

Med moderne sekventeringsteknologi er det muligt at identificere en lang række af mutationer med en meget høj grad af præcision – faktisk kan man med den rigtige tilgang opdage én enkelt mutation i en ud af tusind tumorceller - men der er stadig udfordringer forbundet med metoden. Sekventeringsmaskiner kan forårsage unøjagtigheder i sekvenseringsdata, f.eks. i de tilfælde hvor kvaliten af tumor-materialet ikke er optimal. Da det ikke altid er muligt at ekstrahere DNA og RNA af høj kvalitet fra en tumor prøve.

For at kunne udnytte alle de sekventerede data til behandling, må de underkastes analyse. Der er omkring 6 milliarder bogstaver (nukleotider), der skal læses fra hvert enkelt genom, og at finde netop de sygdomsfremkaldende mutationer i alle disse nukleotider er en udfordring. Et af de mest centrale analyseelementer i forhold til kræftpræcisionsmedicin er de såkaldte "Variants of uncertain significance" (VUS). Begrebet dækker over mutationer, hvis potentielle *driver* effekt og eventuelle potentiale som lægemiddelmål, forskerne endnu ikke kender. I artiklen om kræftbiologi introducerede vi begreberne *driver*- og *passenger*-mutationer og gener i kræftbiologien, f.eks. *TP53*, *EGFR* og *RB1*. Som beskrevet i artiklen, er størstedelen af de kræftmutationer, der findes i en tumor passenger-mutationer (herunder dem i *TP53*, *EGFR* og *RB1*), og det er en helt central opgave at fastslå, hvilke der er driver-mutationer og hvilke der er passenger-mutationer. Hver patients tumor vil indeholde et sted mellem flere hundrede og flere tusinder mutationer, og nogle af disse vil befinde sig i gener som i princippet kan rammes af specifikke lægemidler. For en afdeling der tilbyder præcisionsmedicin, er det afgørende at fastslå, om en bestemt mutation gør kræftcellen potentielt sårbar overfor en målrettet behandling. Det er vigtigt at huske på, at standardbehandling for de fleste kræftformer stadigvæk består af kirurgi, strålebehandling og kemoterapi, men præcisionsmedicinske tilgange bliver stadig mere udbredt, og med den nuværende udvikling, er det realistisk at forvente, at målrettede behandlinger inden for en kort årrække vil være standardbehandling for mange kræftformer.

Målrettede behandlingsformer

Målrettede behandlingsformer er dem, der har en specifik virkning på et aspekt af kræftsygdommen, således at sygdommens vækst, overlevelse eller spredning hæmmes. Disse behandlingsformer udgør grundlaget for præcisionsmedicin, og kan opdeles i to grupper af lægemidler (figur 5):

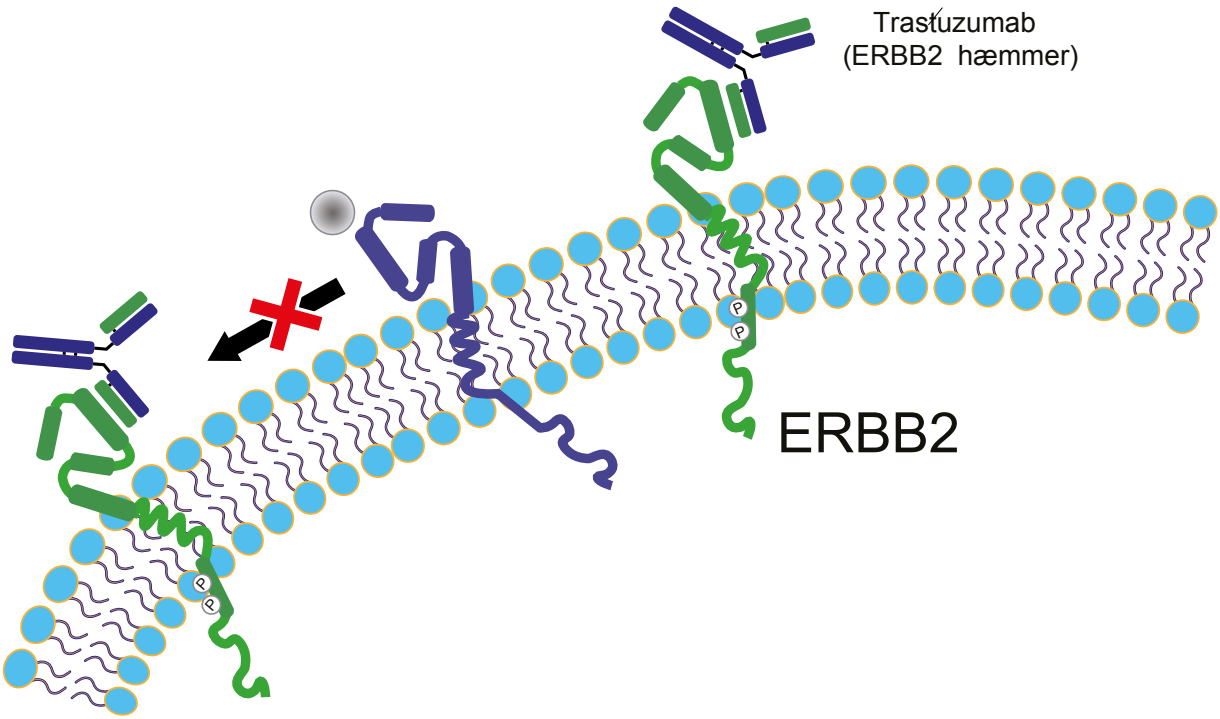


Figur 5. To hovedtyper af præcisionsbehandling/præcisionsmedicin.

Små-molekylelægemidler (til venstre) - er kemikalier, som er i stand til at gennemtrænge cellemembranen uden brug af membranreceptorer eller andre hjælpemekanismer (lipidbaserede bærere). Små-molekylærelægemidler kan potentielt målrettes imod enhver type cellekomponent (organeller, proteiner, DNA, RNA). Antistofbaseret behandling (til højre)- er baseret på såkaldt monoklonale antistoffer (identiske antistoffer som er fremstillet af en klon af identiske immunceller, som bliver ført frem til kræftcellen. Behandlingsformen virker ved enten at introducere antistoffer, som i sig selv er målrettet imod de kræftassocierede proteiner (såkaldt "nøgne antistoffer") eller ved at sammenføje radioaktive stoffer, toksiner eller dræberceller med antistoffet, der specifikt binder til en kræftcelle.

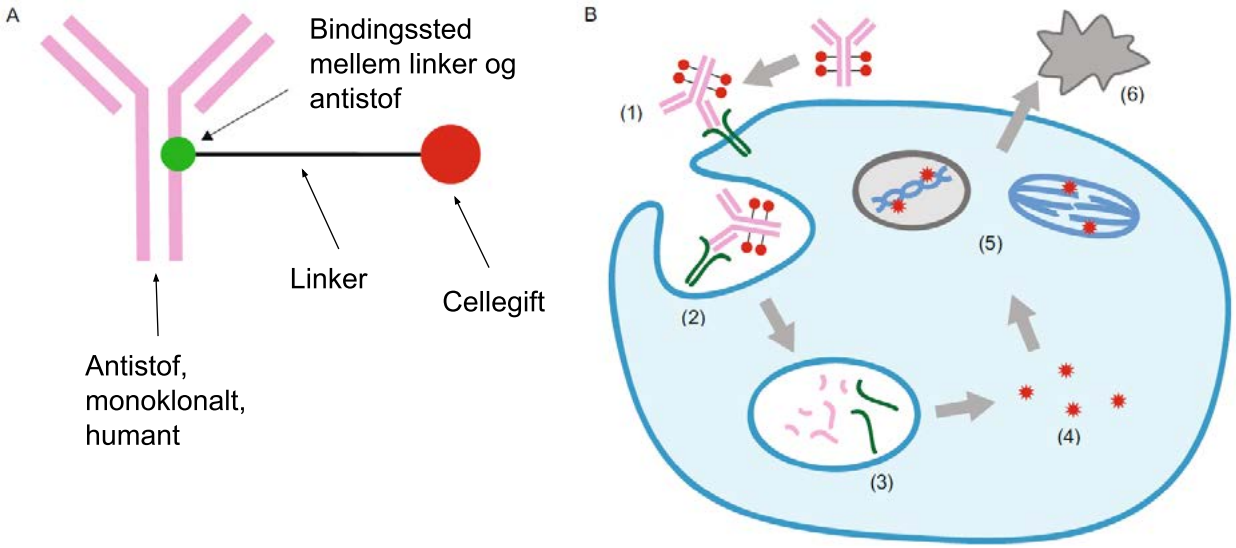
Småmolekylære lægemidler: I modsætning til konventionelle lægemidler, som hovedsagelig er målrettet celleoverfladen, er disse molekyler så små, at de kan trænge ind i cellen og interagere direkte og specifikt med deres mål. Mange af disse småmolekylære lægemidler er designet til at hæmme kinaseproteiner, som er afgørende for at aktivere alle signalkaskader i cellen (f.eks. tyrosin-kinase hæmmer i BCR-ABL1, figur 3). For eksempel er MAP-kinase signalkaskaden (illustreret i artiklen om kræftbiologi), kendt for at være overaktiveret i mange typer af kræft. Adskillige små molekyle-inhibitorer som er målrettede mod proteiner i MAP-kinase signalkaskaden (herunder MEK- og RAF-onkogenerne) er allerede godkendt af sundhedsmyndighederne, og mange flere er under udvikling. Der findes ligeledes godkendte småmolekyle hæmmere, der retter sig mod flere kinaser, der fungerer som receptorer på cellens overflade, såsom EGFR og ERBB2. Disse inhibitorer kan være gavnlige for patienter med aktiverende mutationer eller amplifikationer af disse gener.

Antistofbaseret behandling: Denne type lægemidler udnytter immunsystemets evne til at lokalisere og neutralisere ydre trusler, såsom vira og bakterier. Formålet med behandlingerne er den samme som ved vacciner, nemlig at stimulere immunsystemet til at angribe de syge celler. Antistofbaserede behandlinger er blandt de mest populære (og dyreste) lægemidler i dag på grund af deres specifikke virkning og evne til at ramme ændrede eller stærkt overudtrykte proteiner. Eksempler på succesfulde antistofbaserede behandlinger er hæmmere af ERBB2 (figur 6).



Figur 6. Virkning af antistofpræparatet Herceptin (aktivstof: Trastuzumab) Herceptin virker ved at binde sig til receptoren ERBB2 (der også kaldes HER2). Derved kan vækstfaktor og den anden receptor ikke bindes til ERBB2 og signalkaskaden som vises i artiklen om kræftbiologi kan ikke aktiveres. Således stoppes overaktivering af celledelingen. Samtidig tiltrækker det bundne antistof, Herceptin, immunceller, så kræftcellerne dræbes.

Antistoffer optages typisk af de celler, de binder til, og dette anvendes i såkaldte antistof-cellegift konjugat metoder. Meget kraftige cellegifte er bundet til antistofferne, som dræber de celler som antistoffet binder til



Figur 7. Struktur og virkemåde af antistofkoblet cellegift. (A) Strukturen består af et antistof (monoklonalt, humant), en kemisk linker (forbinder) og en cellegift. Linkeren er forbundet med en kovalent binding til antistoffet. (B) Generel virkemåde for antistofkoblet cellegift. 1) antistofdelen bindes til et overfladeantigen hvilket fører til 2) endocytose af antistof-antigen komplekset. 3) Komplekset optages i et lysosom og enzymerne i lysosomet nedbryder komplekset. 4) Nedbrydningen i lysosomet fører til frigivelse af cellegiften. 5) Cellegiften bindes til DNA og 6) dette fører til celledød. Omtegnet efter Tsuchikama, K. & An, Z. Protein Cell (2018) 9: 33. https://doi.org/10.1007/s13238-016-0323-0, oprindelig licens https://creativecommons.org/licenses/by/4.0/

Som beskrevet i artiklen om kræftbiologi vil kræftceller ofte forsøge at undertrykke kroppens egen immunreaktion over for kræftcellerne. Immunterapi er en meget vellykket type præcisionsmedicin, hvor specifikke molekyler (ofte antistoffer) bruges til at fjerne "bremserne" på kroppens eget immunsystem.

Andre former for kræftbehandling, såsom hormonbehandling og stamcelletransplantation, vil ikke blive beskrevet yderligere her, men disse kan også anvendes præcisionsmedicinsk med udgangspunkt i patientens molekylære sygdomsprofil. I dag er næsten 300 små molekyler og 60 antistofbaserede lægemidler godkendt til behandling i Europa (www.drugbank.ca; www.biopharma.com).

Et patientforløb ved æggestokkræft

I det følgende kan du se et eksempel på, hvordan dataanalyse og præcisionsbehandling kan indgå i forskellige faser af et patientforløb. Patientcasen er opdigtet, men følger guidelines for kræftbehandling.

En kvinde henvender sig til sin praktiserende læge (figur 1), da hun gennem et stykke tid ikke har været helt frisk. Hun har været træt, har ikke haft appetit, og har haft tyngde i benene. Hun har desuden haft lettere smerter i underlivet.

Den praktiserende læge vurderer, at der bør tages en blodprøve, for at undersøge niveauet af proteinet CA125 som er en biomarkør som anvendes for at vurdere en risiko for eventuel æggestokkræft. CA125 kan også ses forhøjet ved godartet sygdom som f.eks. infektion. Resultatet af analysen viser 50 enheder per mililiter (grænseværdien er 35). Lægen henviser derfor patienten til en speciallæge, for at der kan foretages ultralydsscanning og gynækologisk undersøgelse.

Scanningen viser en udfyldning i det lille bækken, og patienten henvises derfor straks til hospitalet med diagnosen *obs. ovarie cancer* (æggestokkræft) i henhold til kræftpakkeforløb.

På hospitalet tages der blodprøver og der foretages gynækologisk undersøgelse samt ultralydsscanning. Resultaterne af disse kombineres i et indeks kaldet Risk of Malignancy Index (RMI). RMI er for denne kvinde 500 (grænseværdien er 200), og patientens behandlingsforløb diskuteres af et multidisplinært team bestående af læger, sygeplejersker og molekylærbiologer med henblik på operation eller kemoterapi.

Operation skønnes mulig med forventet bortoperation af al synlig tumorvæv. Al makroskopisk tumorvæv fjernes ved operationen, og tumorvævet sendes til den patologiske afdeling til undersøgelse. Den patologiske vurdering viser, at der er tale om et stadium IIIC ovarie cancer (æggestokkræft) af typen "high grade" serøst adenocarcinom.

Efter operationen henvises patienten til behandling med Carboplatin og paclitaxel (kemoterapi), som er en evidensbaseret behandling (d.v.s. medicin, som i kliniske forsøg med patienter har vist at virke til denne form for kræft). CA125 niveauet følges og falder efter første kemobehandling til 25 U/ML, hvorefter det stabiliseres på 10 U/ML. Patienten har god effekt af behandlingen.

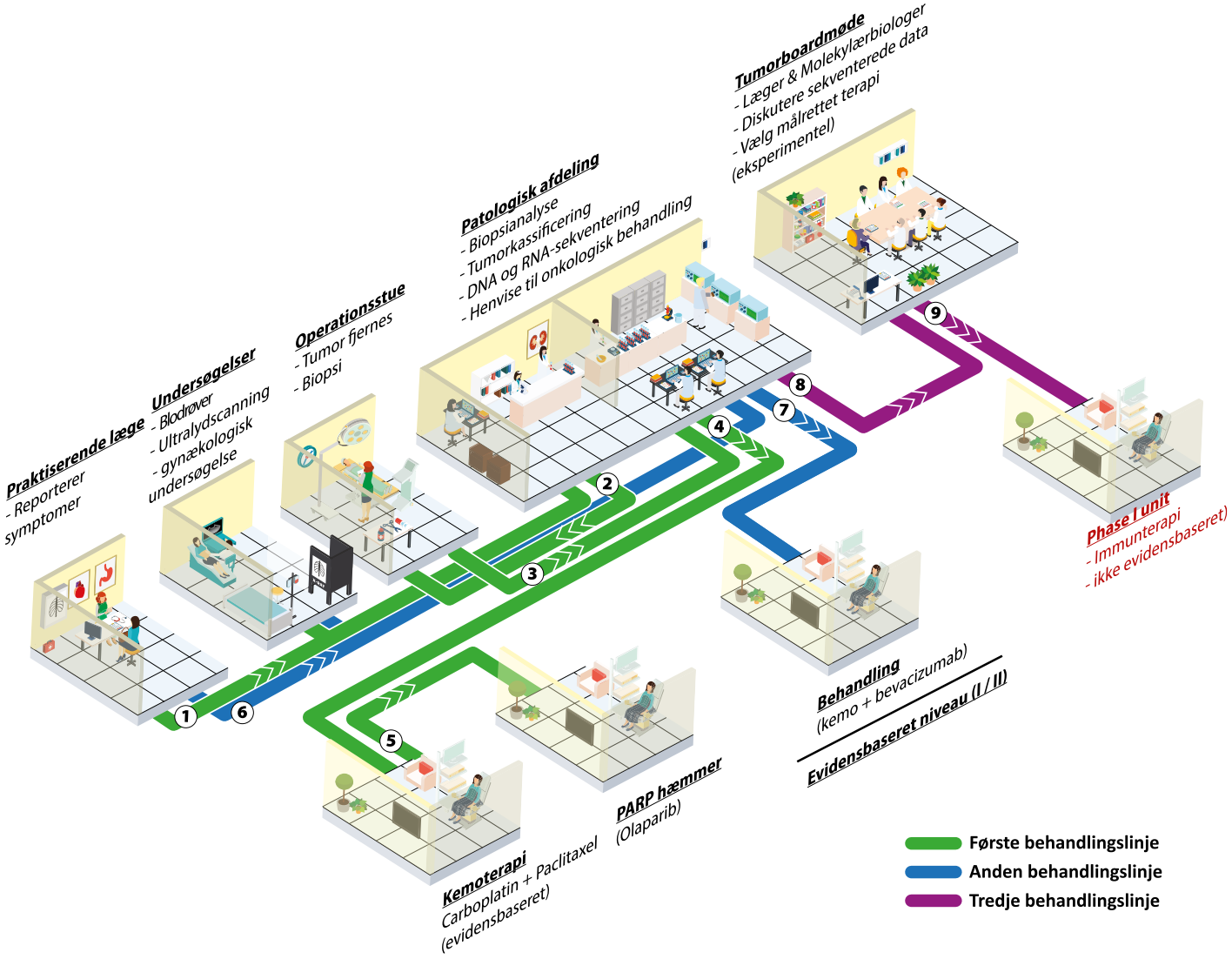
Vævet som er fjernet ved operationen sekventeres og viser, at der er mutationer i genet *BRCA2*. Fordi patienten har denne type mutation, er af typen serøst adenocarcinom og samtidig responderer på kemoterapi med platin, tilbyder man behandling med poly (ADP-ribose) polymerase (PARP) hæmmer (Olaparib). Det hæmmer virkningen af PARP, som er et enzym, som bidrager til reparation af enkeltstrengsbrud i DNA. Når DNA ikke kan repareres af denne grund, dør kræftcellerne, om end det kun gælder nogle typer af kræftceller, herunder nogle, hvor *BRCA2* er muteret. Kvinden modtager behandling med PARP hæmmer.

Efter 14 måneder henvender kvinden sig igen grundet symptomer på tilbagefald af sygdom (tyngdefornemmelse i underlivet). Der foretages gynækologisk undersøgelse samt scanning og der tages en blodprøve. Blodprøven viser, at CA125 niveauet er steget til 450 U/ML og scanningen viser tumorvækst. Patienten har tilbagefald af sygdom. Da kvinden har haft god effekt af tidligere medicinsk behandling tilbydes yderligere kemoterapi samt deltagelse i klinisk forsøg.

Efter 6 måneder kan lægerne konstatere, at kvindens tumor er vokset. Der er ikke umiddelbart flere muligheder for anden behandling (andre typer kemoterapi/kliniske forsøg). Patienten er forsat oppegående og deltager i sociale arrangementer og henvises derfor til en særlig onkologisk afdeling, med henblik på muligheden for eksperimentel behandling. Ved henvisning undersøges patientens tumorvæv (ofte en ny biopsi) med store genetiske undersøgelser for at afklare og tumorprofiler giver mulighed for behandling efter denne med lægemidler som er godkendt til andre kræftsygdomme. Eksperimentel betegner behandling som udføres som del af et videnskabeligt forsøg med flere deltagere. Det kan dog også være ikke-afprøvet behandling, der sammensættes til en enkelt patient, fordi der ikke er andre behandlingsmuligheder.

Resultatet af sekventeringen viser den tidligere beskrevne *BRCA* mutation, og finder desuden mutationer i et meget stort antal gener. Dvs. at der er højt tumormutationsload (antal mutationer pr megabase), og patienten tilbydes derfor immunterapi.

Et patientforløb ved æggestokkræft



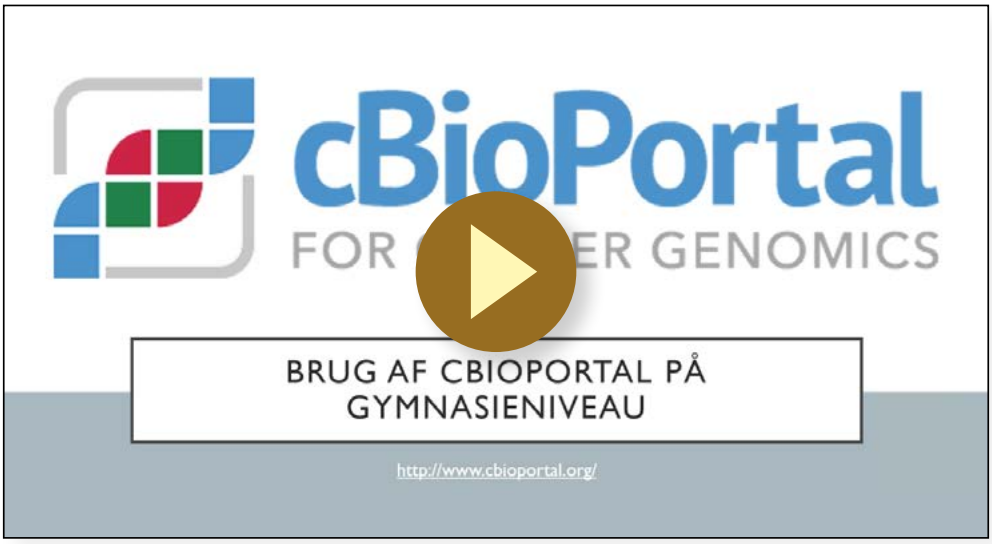
1. Besøg hos egen læge: Rapportering af symptomer. Patienten henvises til speciallæge der foretager Undersøgelser: Blodprøver, ultralydsscanning, gynækologisk undersøgelse.
2. Hospitalsundersøgelser, diagnose, henvisning til kirurgi.
3. Operationsstue: Tumor fjernes og sendes til biopsiundersøgelse.
4. Patologisk enhed: analyse af biopsi, klassificering af tumor, DNA og RNA sekventering. Patologivurdering viser at det drejer sig om et stadium IIIC "high grade" æggstokkræft. Patienten henvises til onkologisk behandling med Carboplatin og Paclitaxel (evidensbaseret).
5. Sekventeringsdata viser mutation på *BRCA2* gen. Patienten behandles med PARP hæmmer (olaparib).
6. 14 mdr. senere: undersøgelser viser tilbagefald.
7. Behandlingsrunde 2**:** platinbaseret kemoterapi kombineret med (klinisk forsøg).
8. Der er ingen tilfredsstillende respons på behandlingen. Patienten henvises til fase 1 enheden.Tumorceller undersøges ved sekventering med store paneler som indeholder gener/hotspots for alle kendte behandlingstargets.
9. Sekventeringen viser den tidligere beskrevne BRCA mutation, og finder desuden mutationer i et meget stort antal gener. Patienten tilbydes immunterapi.

Bemærk:

1. Grønne, blå og violette spor repræsenterer henholdsvis først, anden og trejde behandlingslinje.
2. Pilene på farvesporerne repræsenterer behandlingsprocedurens retning.

Introduktion til cBioPortal

I denne screencast får du en generel introduktion til databasen cBioPortal, som du skal arbejde med i biotekøvelserne.



Biotekøvelse 1. Brystkræft og genet *ERBB2*: Ændringer i transskription og translation.

Formål

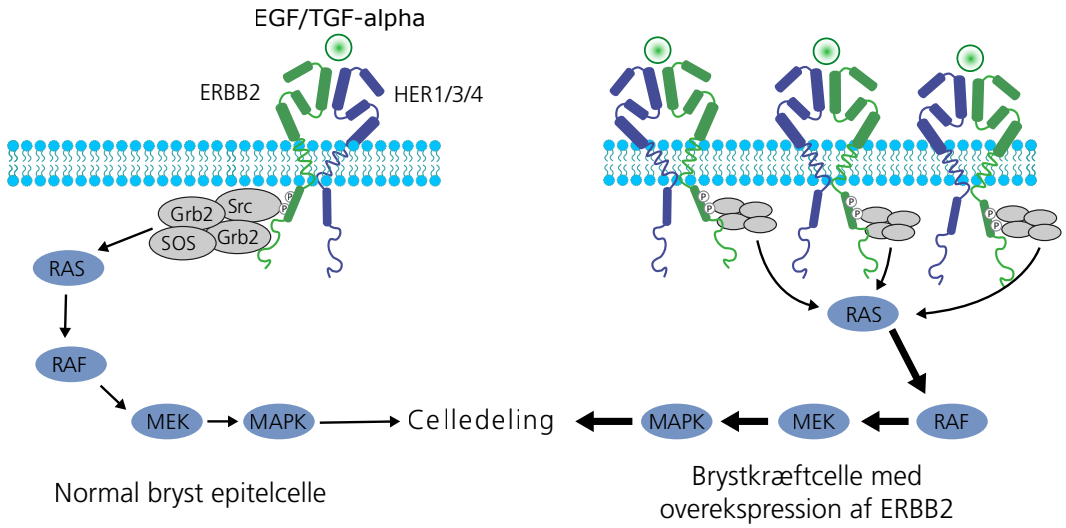
- At undersøge forholdet mellem kopi-antallet af genet *ERBB2* og niveauet af mRNA (dvs. ændringer i genekspression på transskriptionsniveau)
- At undersøge sammenhængen mellem niveauet af mRNA for genet *ERBB2* og niveauet af proteinet, som genet koder for (dvs. ændringer i genekspression på translationsniveau)

Baggrund

I Danmark diagnosticeres årligt ca. 4.650 nye tilfælde af brystkræft. Når tumorerne er fjernet ved operation, bliver de undersøgt ved hjælp af forskellige molekylærbiologiske metoder. Hermed kan lægerne påvise ændringer i patienternes gener, og ud fra disse resultater træffe valg om hvilken behandling, patienten skal tilbydes.

Nogle brystkræftpatienter har ændringer i genet *ERBB2*, som koder for en receptor, ERBB2[1], der sidder i cellemembranen. Af historiske årsager kaldes receptoren også for HER2, men ERBB2 er det korrekte proteinnavn. I normale celler aktiveres denne receptor ved binding af vækstfaktorer fra andre celler, og gennem en længere signalvej fører bindingen af vækstfaktorer til stimulering af celledeling, se figur 1, venstre side.

[1] Gener skrives med store bogstaver og skrå skrift, f. eks. *ERBB2*. Proteiner skrives med almindelig skrift, f. eks. ERBB2.



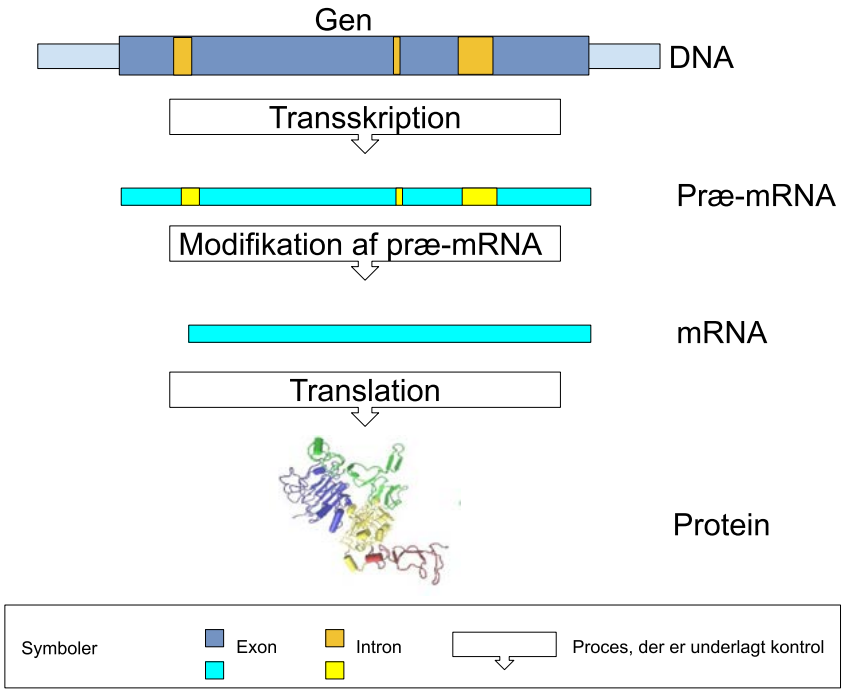
Figur 1. Virkemåde for ERBB2-receptoren i normale brystceller (til venstre) og i visse brystkræftceller (til højre).
Venstre side af figuren: I normale brystceller forbindes receptoren ERBB2 med andre receptorer og dette kompleks virker tilsammen som receptor for vækstfaktoren EGF eller vækstfaktoren TGF-α. Når receptoren aktiveres, udløses en signalkaskade som stimulerer celledeling. **Højre side af figuren:** I brystkræftceller hos nogle patienter findes der amplifikationer af *ERBB2* genet, som koder for receptoren ERBB2, hvilket fører til, at der er for meget ERBB2 i cellemembranen. Når der er ekstra af denne receptor, stimuleres en for stor signalering videre i signalvejene, hvilket i den sidste ende fører til abnorm celledeling.

I tumorer fra nogle af brystkræftpatienterne er der amplifikationer af genet *ERBB2*. Amplifikation betyder, at der findes betydeligt flere kopier end de normale to af genet. Der er også en overekspression af det protein, som genet koder for, så der er et forøget niveau af receptoren ERBB2 i cellemembranen (figur 1, højre side). Det er kendt, at patienter med *ERBB2* amplifikation har en forholdsvis høj risiko for at sygdommen vender tilbage efter operation, og behandlingen skal indrettes derefter.

Hvis man vil forstå mekanismerne bag udvikling af kræft, er det ikke nok udelukkende at undersøge ændringer i cellernes proteinniveauer eller af deres DNA. Alle elementer af proteinsyntesen (fra DNA til præ-mRNA og videre til mRNA og protein) og reguleringen af disse elementer (se figur 2), kan være relevante at undersøge. For eksempel er det i brystkræft vigtigt at forstå, hvordan forskellige ændringer i det førnævnte gen *ERBB2* påvirker både transskription (syntese af mRNA) og translation (mRNA-kodens oversættelse, som fører til syntese af protein).

Hvis forholdet mellem niveauerne af mRNA og protein er stabilt, er det også interessant ud fra et forskningsmæssigt perspektiv. Undertiden har man kun en værdi for enten mRNA eller protein fra en bestemt patient, og man kan have brug for den anden værdi i en forskningsmæssig sammenhæng. Derfor kan det være af interesse at kunne finde frem til en omregningsfaktor fra den ene værdi til den anden værdi.

I denne opgave skal du undersøge, hvordan niveauerne af mRNA og af protein påvirkes af amplifikationer i genet *ERBB2* hos brystkræftpatienter.

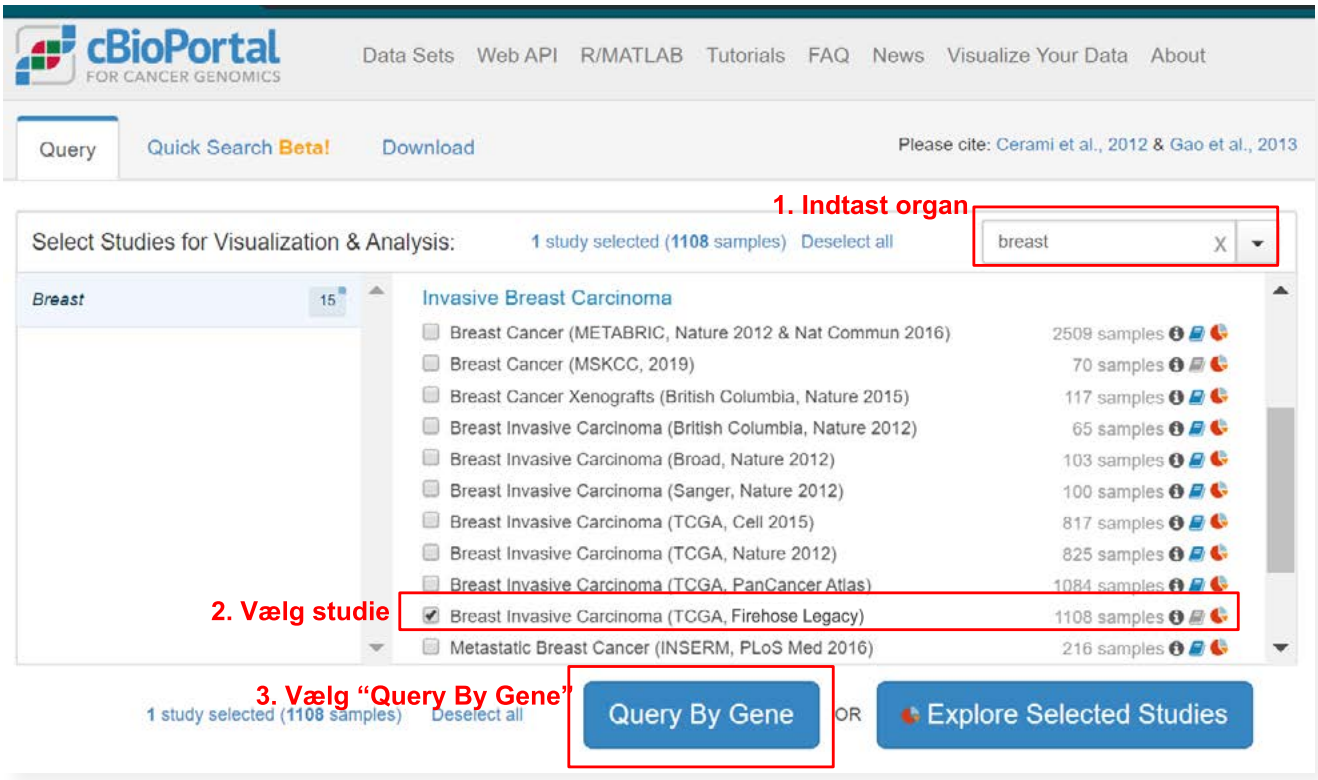


Figur 2. Proteinsyntesen med nogle af de delprocesser, der bliver reguleret af cellen og/eller omgivelserne.

Fremgangsmåde og arbejdsspørgsmål.

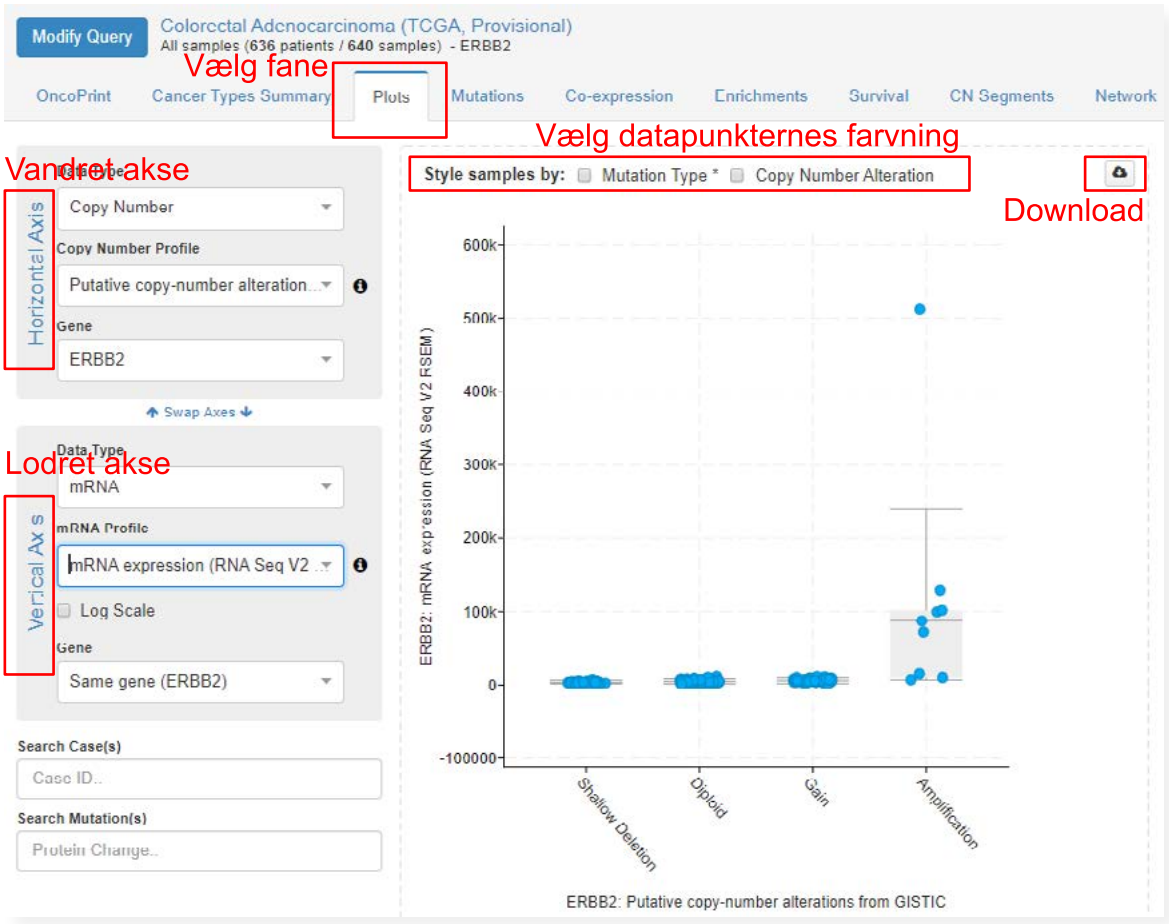
Der er i alt 17 punkter i fremgangsmåden, og der er spørgsmål fra A-I.

1. Gå til [cBioPortal's hjemmeside.]
   I feltet "Search" tastes organet som man vil arbejde med, her "breast" (figur 3).



Figur 3. Indtast organ i det felt på hjemmesiden, hvor der til at starte med står "Search".
Sæt hak ved det ønskede studie og vælg "Query By Gene".

2. Under "Invasive Breast Carcinoma" (som er en variant af brystkræft), sæt hak ved "Breast Invasive Carcinoma (TCGA, Firehose Legacy)" (figur 3). Nu har du valgt en bestemt, stor undersøgelse, som er udført med netop denne type af brystkræft.
3. Vælg "Query By Gene" (figur 3).
4. Find "Enter genes"-boksen. I underboksen i midten, hvor der står "Enter HUGO gene symbols, Gene Aliases, or OQL" skal du taste gennavnet *ERBB2* (med almindelig skrift, ikke skrå skrift). Tryk på "Submit Query".
5. Efter lidt ventetid kommer der et resultatvindue frem, som starter på fanen "OncoPrint".
6. Skift til fanen "Plots". Se figur 4 for et eksempel *fra et andet studie*.



Figur 4. Fanen "Plots". figuren viser hvor man finder fanen "plots" og hvordan man skal indstille til at vise mRNA-niveau som funktion af antallet af kopier af genet *ERBB2*. Plottet er fra en anden kræftform end den, som du selv skal arbejde med, så derfor ser det ikke ud som dit plot.

Undersøg forholdet mellem ændringer i genet og niveau af mRNA

7. Efter at der er skiftet til fanen "Plots", skal det vælges, hvordan plottet skal se ud. På figur 4 vises, hvordan der skal indstilles, og det forklares også i de næste punkter.
8. På menuen for den vandrette (*Horizontal*) akse vælges "Copy Number" og "Putative Copy Number Alterations from GISTIC"
9. På menuen for den lodrette (*Vertical*) akse vælges "mRNA" og derefter nøjagtigt denne tekst "mRNA expression (RNAseq V2 RSEM)" (vær især opmærksom på hvad der står efter "mRNA expression").
10. Ved datapunkternes farve skal der IKKE være valgt noget. Der skal IKKE være hak ved Log Scale ved den lodrette akse.
11. Download figuren fra cBioPortal. Behold vinduet åbent, for det skal bruges igen efter, at du har besvaret nogle arbejdsspørgsmål.

12. Læs begrebsforklaringer (figur 5) til den downloadede figur og besvar derefter de nedenstående arbejdsspørgsmål ud fra den downloadede figur:

Begrebsforklaringer til figuren hvor mRNA vises for forskellige antal kopier af *ERBB2:*

Deep deletion: Hele genet er forsvundet på begge kromosomer

Shallow deletion: En del af genet mangler på en eller begge kopier af genet

Diploid: Det almindelige antal af gener, nemlig to gener, et på hvert kromosom

Gain: Der er *relativt få* ekstra kopier af genet

Amplification: Der er *relativt mange* ekstra kopier af genet

Figur 5. Begrebsforklaringer.

Arbejdsopgaver:

A. Hvilken type af ændring i genet *ERBB2* i brystkræfttumorer ser ud til at give en øget transskription (øget dannelse af mRNA)?

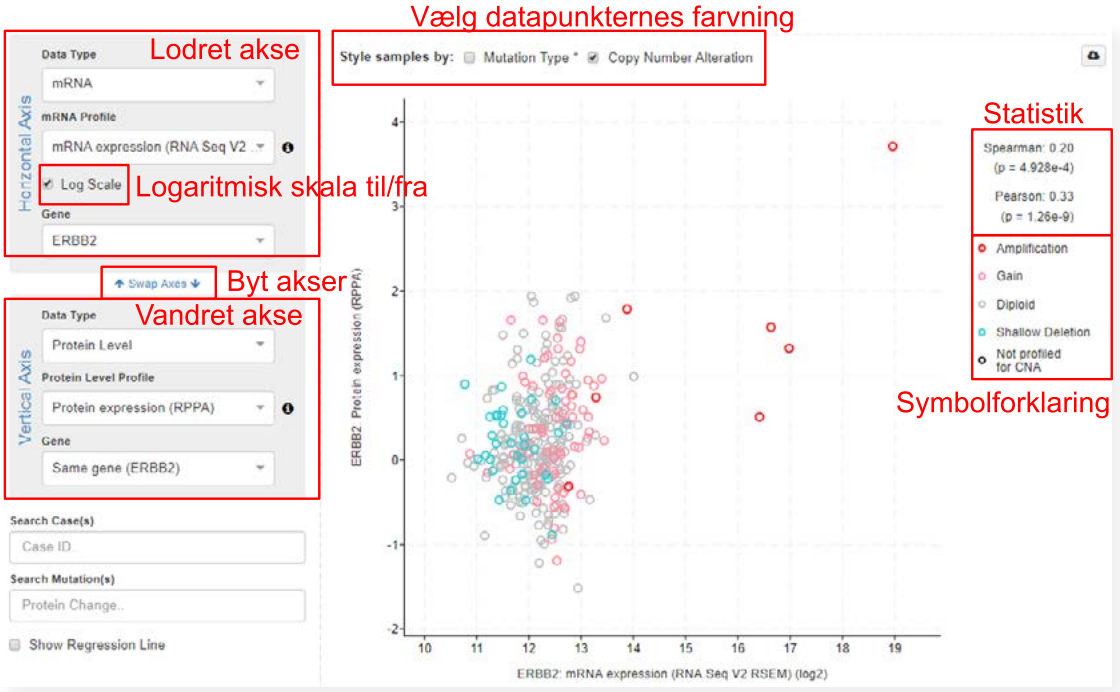
B. Læs begrebsforklaringen til denne type ændring i ovenstående liste.

C. Tænk nu på de forskellige led i proteinsyntesen og forklar med brug af fagudtryk, om den øgede transskription i forbindelse med de fundne typer af ændringer giver mening.

D. Opstil en fagligt begrundet hypotese om, hvad der sker med niveauet af protein, hvis niveauet af **mRNA** for genet *ERBB2* øges? Følg derefter de følgende trin (13-17) for at få data til at undersøge hypotesen.

Undersøg forholdet mellem niveauet af mRNA og niveauet af protein, for genet *EBRR2*

Gå tilbage til søgningen i *cBioPortal* fra før. Det foregående plot skulle gerne være der stadig. Fortsæt sådan (se også figur 6):



Figur 6. Indstillinger til plot af proteinniveau mod mRNA-niveau *fra en anden kræftform* end den, som du arbejder med. Se instruktioner i teksten.

13. Tryk én gang på tasten "Swap axes" for at skifte mellem akserne.
14. Sæt hak ved "Log scale".
15. Vælg i menuen for den lodrette akse: "Protein level" og "Protein expression (RPPA)"
16. Under "Datapunkternes farve" vælges at datapunkterne skal farves efter ændringer i antal kopier af genet (markér ved Style Samples by Copy Number).
17. Den fremkomne graf viser niveauet af det protein, som *ERBB2* koder for, plottet mod niveauet af mRNA for samme gen. Lav et screenshot eller download plottet. Besvar arbejdsspørgsmål E:

E. Hvordan stemmer data overens med hypotesen fra punkt D – ud fra et *umiddelbart skøn*? Undersøg især hvilken form for sammenhæng (ingen, lineær, eksponentiel eller andet) der ser ud til at være, og om sammenhængen ser ud til at være stigende eller faldende. Det er med til at afgøre, om det overhovedet giver mening at gå i gang med en statistisk analyse, og hvilken analyse, der skal bruges.

Næste trin: statistisk analyse. Har klassen matematik med i forløbet?

Ja. Så skal du gøre dig klar til matematik ved at læse teksten "Fra bioteknologisk analyse til statistisk analyse". Derefter skal du lave arbejdet i matematiktimerne (Matematikøvelse 1). Når du har formuleret en konklusion på den statistiske analyse af figurens data og din hypotese i punkt D og E med både matematisk og bioteknologisk sprog, skal du gå videre i punkt F-H.

Nej. Så skal du læse teksten "Pearsons test". Når du har anvendt oplysningerne derfra på din figur og din hypotese i punkt D og E, skal du gå videre i punkt F-H. Bemærk, at den statistiske analyse i cBioportal kun giver mening hvis en række antagelser er overholdt, herunder: Punkterne skal tilnærmelsesvis ligge på en ret linje og der må ikke være punkter som ligger anderledes. Målingerne skal være uafhængige, dvs. at en måling ikke siger noget om den næste (patienter må f.eks. ikke være i familie).

Læs, når du IKKE har matematik med

Pearsons test

I denne tekst henvises til statistikboksen i eksemplet i figur 6, som er fra en anden undersøgelse end din egen. Du skal bruge det nedenstående til at fortolke din egen undersøgelse og dine hypoteser fra punkt E og F.

Hvad testes i forbindelse med Pearsons korrelation?

Det som testes i cBioPortal (figur 6), er en nulhypotese om, at der *ikke* er en lineær korrelation mellem niveauet af mRNA og protein i patienternes tumorer. Det er formodentlig det stik modsatte af, hvad du har formuleret i punkt F, men den statistiske test fungerer altså ved at lave den nævnte nulhypotese.

Hvordan bruges *p*-værdien?

P-værdien i forbindelse med Pearsons korrelation i statistikdelen af resultatvinduet angiver en talværdi, som bruges til at afgøre, om nulhypotesen skal forkastes.

Ofte bruger man i bioteknologi og mange andre fag, at *p*-værdien skal være under 0,05, for at man kan konkludere, at nulhypotesen skal forkastes. I eksemplet i figur 6 er *p*-værdien angivet til 1.26e-9, hvilket skal læses som $1.26 \cdot 10^{-9}$, hvilket er langt under 0,05. Derfor skal nulhypotesen i det eksempel forkastes, og man må konkludere, at der er en lineær korre-

udelukke at der for eksempel er to eller flere søstre, der begge har brystkræft, som indgår i data. Men på den anden side er der så mange datapunkter, at nogle få uafhængige data næppe vil kompromittere hele analysen.

Nu skal du lave matematikken og derefter gå til punkt G-I.

lation mellem niveauerne af protein og mRNA i patienternes tumorer i eksemplets studie.

Hvordan bruges Pearsons korrelationskoefficient?

Tallet lige efter ordet "Pearson", 0.33, er korrelationskoefficienten og tolkes som følger: Tallet er et slags udtryk for, hvor tæt punkterne ligger på en ret linje.

Tallet vil ligge mellem -1 og 1, og være negativ for lineært faldende sammenhæng og positivt ved en lineært voksende sammenhæng.

Hvis tallet er 0 er der ingen sammenhæng, og hvis tallet er henholdsvis -1 eller 1 er der den bedst mulige sammenhæng (alle datapunkter ligge på en ret linje). Det er ikke muligt at give en absolut værdi til at afgøre om sammenhængen er god eller dårlig, så dette tal kommer du ikke til at anvende i din analyse, idet det kræver en del matematiske overvejelser og dermed at en matematiklærer deltager.

Hvilke data indgår i analysen?

I analysen i cBioPortal indgår data fra patienter med mange forskellige ændringer i *ERBB2*, ikke kun amplifikationer. Det giver en uklarhed i tolkningen, men når ikke der er matematik med, er det desværre ikke muligt at få lavet en isoleret analyse kun med patienter med amplifikationer.

Nu skal du bruge det ovenstående når du går videre til punkt G-I.

Arbejdsopgaver:

F. Hvorfor er det vigtigt at udføre statistiske tests, før man kan drage videnskabeligt velunderbyggede konklusioner?

G. Hvilke konklusioner kan du nu, formuleret på bioteknologi-sprog, drage om din hypotese fra punkt D og E på baggrund af de udførte analyser og statistiske tests?

H. Hvad sker der med styringen af cellevæksten, hvis der er et øget niveau af det protein, som *ERBB2* koder for? Her skal oplysningerne i afsnittet "Baggrund" inddrages.

Biotekøvelse 2. Hjernekræft: Mutationer i gener i den samme signaleringskaskade

Formål

- at anvende den åbne database *cBioPortal* til at undersøge ændringer i tre gener i den samme signaleringskaskade i hjernetumorer.
- at undersøge om der er mutationer i flere af disse gener i den samme patients tumor, eller om mutationer i det ene gen udelukker ændringer i de andre.
- at diskutere, hvilken betydning disse genændringer har for styring af cellevækst og celledeling via RB-signaleringskaskaden.

Baggrund

Kræft udvikler sig som følge af ukontrolleret cellevækst og -deling. RB-signaleringskaskaden (figur 1) udgør en række af de processer, som har indflydelse på den sunde celles styring af vækst og celledeling. Hvis en eller flere af mekanismerne i kaskaden ikke fungerer korrekt, kan det føre til tab af kontrol over vækst og celledeling. Dog har cellerne også andre kontrolsystemer, som i nogle tilfælde kan træde til og afhjælpe manglerne.

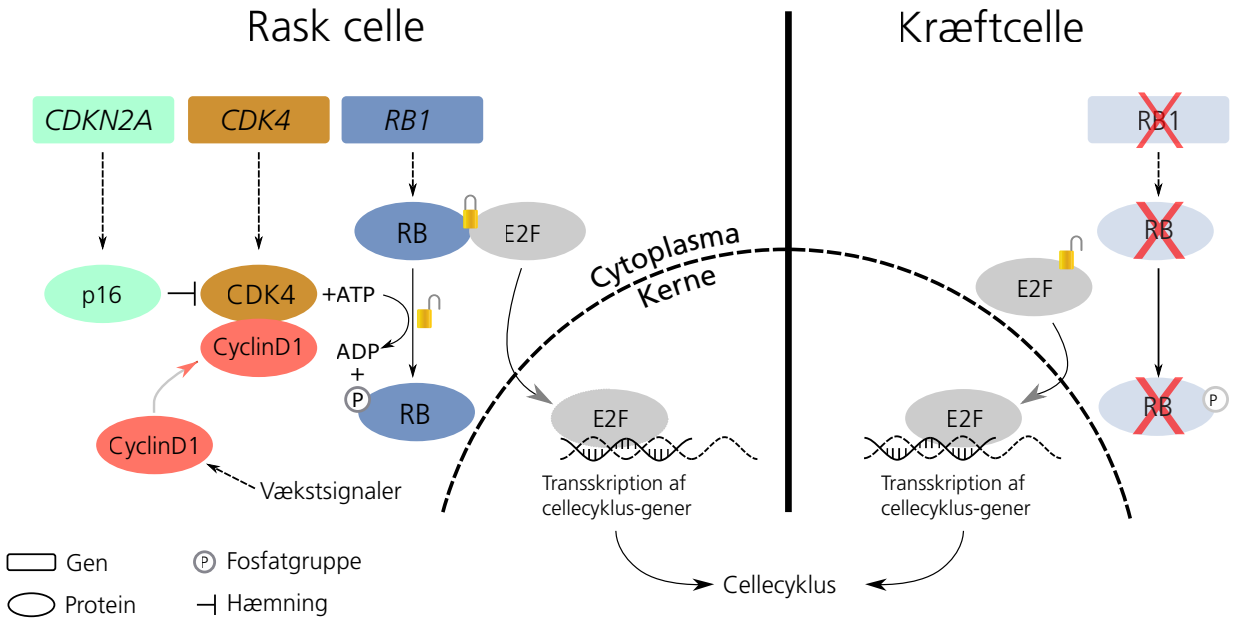
I denne øvelse skal du undersøge resultaterne af et videnskabeligt studie om Glioblastoma, som er den almindeligste og mest aggressive form for hjernekræft. Du skal undersøge ændringerne i tre gener fra RB-signaleringskaskaden i tumorer fra en række patienter.

De tre gener, hvis normale funktion ses på figur 1, er:

- *CDKN2A*, der koder for proteinet p16
- *CDK4*, der koder for proteinet CDK4
- *RB1*, der koder for proteinet RB.

For at forstå de biologiske processer og signaleringsveje, som har betydning i forskellige kræftformer, er det relevant at undersøge, om ændringer i flere specifikke gener hyppigt optræder sammen (*co-occurring*), eller om generne tværtimod sjældent eller aldrig er ændret i den samme tumor (*mutually exclusive*).

I denne øvelse skal du undersøge, om ændringer i generne *CDKN2A*, *CDK4* og *RB1* er *co-occurring* eller *mutually exclusive* i forbindelse med hjernekræft.



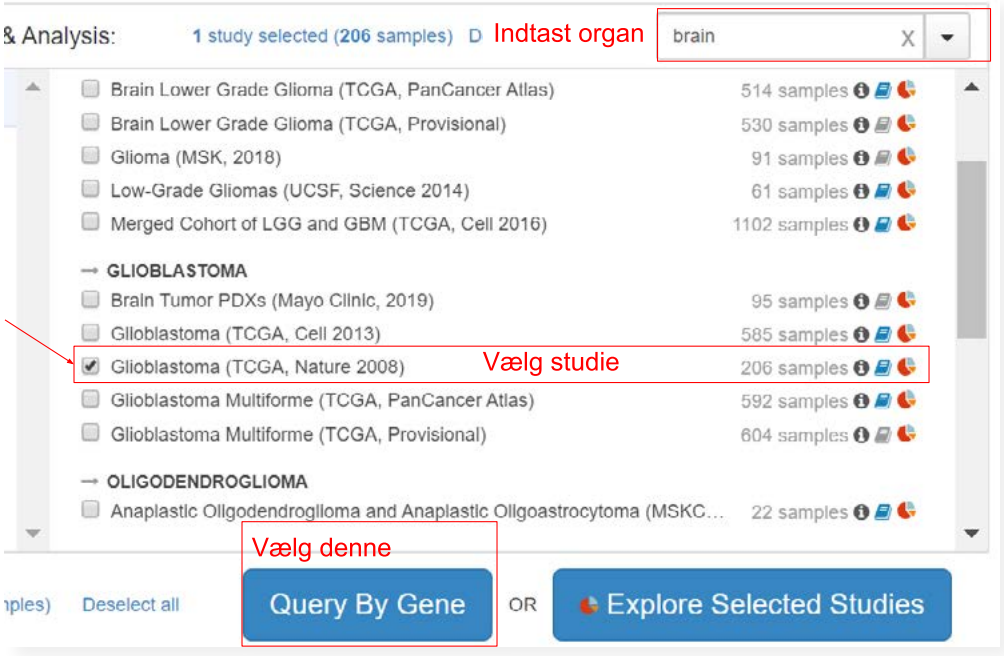
Figur 1. Retinoblastom (RB)-proteinet er en hovedregulator af cellecyklus i den normale celle: Venstre side af figuren: Transskriptionsfaktoren E2F er en aktiverende transskriptionsfaktor for gener, hvis proteinprodukter stimulerer cellecyklus. Når proteinet RB ikke er fosforyleret, hindrer det at cellen går videre i sin cellecyklus. Det sker ved at RB fastholder E2F og derved hindrer E2F i at fremme transskription af gener, der stimulerer cellecyklus. Når cellen modtager et vækstfremmende signal, vil CyclinD1 binde sig til CDK4 og dette kompleks kan fosforylere RB-proteinet. Den fosforylerede udgave af RB kan ikke binde og hæmme E2F, som nu er fri, og derved kan E2F fremme transskriptionen af cellecyklus-generne. Flere proteiner påvirker denne regulering, idet proteinet p16 (udtrykt af genet *CDKN2A*), virker som en bremse på celledelingen i den normale celle. Det sker ved, at p16 bindes til CDK4 og derved hæmmer CDK4's binding med CyclinD1. Når CyclinD1 hæmmes i at blive bundet til CDK4, hæmmes også fosforyleringen af RB, og dermed hæmmes celledelingen. I denne figur ses altså et eksempel på den komplicerede balance mellem fremmende og hæmmende signaler til cellecyklus. Højre side af figuren: Både generne *RB1* og *CDKN2A* er ofte muteret eller helt deleteret (slettet) i kræftceller. figuren viser, hvordan deletion af *RB1* fører til konstant aktivering af cellecyklus.

Fremgangsmåde og arbejdsspørgsmål.

Der er i alt 5 punkter i fremgangsmåden, og der er spørgsmål fra A-I.

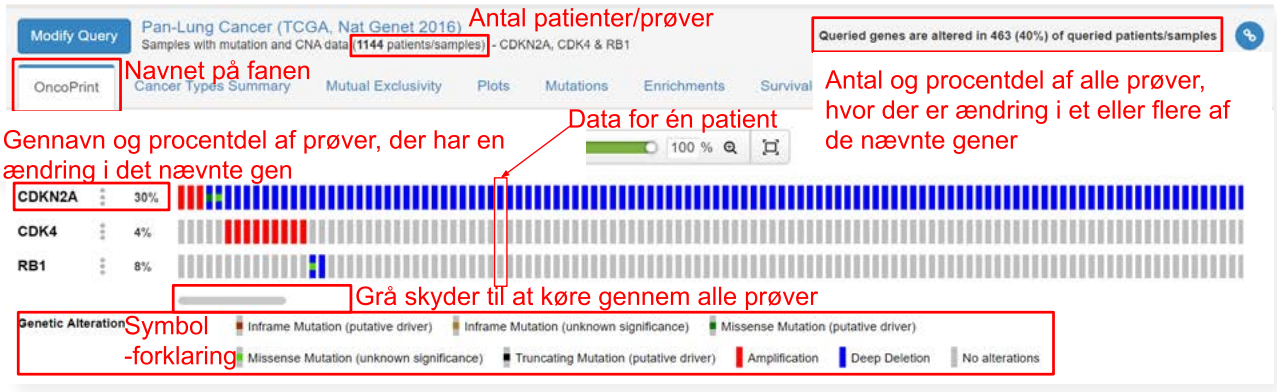
Find datasættet frem

1. Gå til [cBioPortal's hjemmeside](). I søgefeltet tastes det organ, som vil arbejde med, her *brain* (figur 2).



Figur 2. Screenshot, der viser, hvor der skal indtastes organ i søgefeltet, derefter markeres for at vælge studie og endelig vælges søgemetoden "Query by Gene".

2. Under "→Glioblastoma" (dvs. Hjernekræft), vælg "Glioblastoma (TCGA, Nature 2008)". Nu er der valgt en bestemt undersøgelse, som er udført med netop denne type kræft.
3. Vælg søgemetoden "Query By Gene" (se nederst figur 2).
4. I boksen "Enter genes" findes underboksen i midten, hvor der står "Enter HUGO gene symbols, Gene Aliases, or OQL". Her skal tastes gennavnene ***CDKN2A CDK4 RB1***. Tryk på "Submit Query". Vær opmærksom på, at der ved fejl i gennavnene kommer en rødt farvet besked op om, at gennavnene ikke er korrekte, og så kan søgningen ikke startes. Tjek efter om du har skrevet gennavnene helt nøjagtigt.
5. Efter lidt ventetid kommer der et resultatvindue frem, som starter på fanen "OncoPrint". Det ligner figur 3, som dog er fra en anden kræftform. I billedteksten til figur 3 er forklaret, hvordan resultaterne aflæses. Så billedteksten er vigtig at læse. Brug resultaterne fra egen søgning (altså ikke figur 3) til at besvare spørgsmålene i næste punkt.



Figur 3. Resultater fra søgning på de samme gener som skal bruges i denne øvelse, men med en anden type cancer. Billedet viser kun en lille del af undersøgelsens resultater, da der indgår mange patienter i denne undersøgelse. Man skal sikre sig, at man kan se information for alle undersøgte patienter: Hvis der er en grå skyder i billedet, så træk mod højre og venstre for at gennemgå alle patienternes data. Der er markeret hvor på figuren man kan finde de forskellige typer af information. Symbolerne i symbolforklaringen, som er relevante for jeres opgave, er forklaret herunder:

Symbolforklaring :

Amplification/Amplifikation: Der er flere kopier end normalt af dette gen i genomet, dette er en mutation på kromosomniveau.

Deep deletion/Dyb deletion: Genet mangler helt på begge kromosomer, dvs. homozygot for deletionen, en kromosommutation.

Missense mutation: Putative driver/formodet driver) (en genmutation som fører til, at en-flere aminosyrer udskiftes med andre aminosyrer. Med "putative driver" menes, at man har en formodning om, at denne mutation kan være central som årsag til cancers opståen og/eller udvikling.

Missense mutation: Unknown significance/ukendt betydning) (en genmutation som fører til, at en-flere aminosyrer udskiftes med andre aminosyrer. Med "ukendt betydning" menes, at man ikke kender betydningen af denne mutation med hensyn til cancer.

Truncating mutation: En genmutation, som fører til at proteinet, som genet koder for, bliver for kort. Et codon for en aminosyre er blevet erstattet af et stop-codon.

No alterations/ingen ændringer: Der er i studiet ikke fundet ændringer i de nævnte gener i de deltagende patienters tumorer.

Undersøg hvilke ændringer, der er fundet i generne ***CDKN2A***, ***CDK4*** og ***RB1*** i tumorer fra hjernekræftpatienter, og hvilken betydning de har for styring af cellens vækst og deling, gennem besvarelse af arbejdsspørgsmålene herunder.

Besvar følgende spørgsmål ud fra eget søgeresultat fra cBioPortal om de ovennævnte gener inden for Glioblastoma:

A. Hvor mange patienters data indgår i undersøgelsen?

B. Hvor stor en procentdel af patienternes tumorer har en eller anden type af ændring i disse gener:
   - *CDKN2A*
   - *CDK4*
   - *RB1*

C. *RB1*
   - Hvilke(n) af de observerede genetiske ændringer i *RB1*-genet kunne føre til produktion af et ikke-funktionelt protein RB? (En forklaring af de forskellige typer af mutationer kan ses under figur 3).
   - Hvilken konsekvens kan et ikke-funktionelt RB-protein få for forløbet af RB-signaleringskaskaden, og giver det mening i forhold til udvikling af kræft?

D. *CDK4*
   - Hvilke(n) af de observerede genetiske ændringer kunne føre til øget produktion af CDK4-proteinet?
   - Hvilken konsekvens kunne overproduktion af CDK4-proteinet have for forløbet af RB-signaleringskaskaden, og giver det mening i forhold til udvikling af kræft?

E. *CDKN2A*
   - Hvilke(n) af de observerede genetiske ændringer kunne føre til dannelse af et ikke-funktionelt protein p16?
   - Hvilken konsekvens kunne det have i RB-signaleringskaskaden, og giver det mening i forhold til udvikling af kræft?

F. Ser det ud til, når man kigger hen over resultatsiden OncoPrint, at de tre gener er *mutually exclusive* (sjældent eller aldrig er ændret i den samme tumor) eller *co-occurring (*d.v.s. at ændringer i generne som regel optræder sammen*)*? Afkryds nu i nedenstående skema en hypotese for hvert af de viste genpar med hensyn til, om de ser ud til at være *co-occurring* eller *mutually exclusive*.

Gen 1	Gen 2	Skønnet til at være co-occurring (genændringer i begge gener forekommer oftest i samme prøve)	Skønnet til at være mutually exclusive (genændringer i begge gener forekommer sjældent eller aldrig i samme prøve)
CDK4	***RB1***		
CDK4	***CDKN2A***		
RB1	***CDKN2A***		

Næste trin: statistisk analyse.
Har klassen matematik med i forløbet?

Ja. Undersøg med statistiske metoder i matematiktimerne (matematikøvelse 2) hvilken statistisk evidens der er for de hypoteser, du har lavet i punkt F. I matematikøvelsen kommer du til at omdanne din bioteknologiske hypotese til en nulhypotese, der kan testes med matematiske værktøjer. Når du er færdig med matematikøvelse 2, skal du gå videre til punkt G.

Nej. Så skal du bruge den statistiske analyse, som cBioPortal leverer:

- Skift i den søgning, som du var i gang med, til fanen "Mutual Exclusivity". På den kan du se resultaterne af en statistisk analyse, hvor der testes en nulhypotese i retning af, at forekomsten af mutationer i generne er uafhængige (af hinanden). Hvis *p*-værdien ≤0,05, siges testen at være signifikant (på et 5% signifikansniveau).
- Er log2 Odds Ratio større end nul, er der tale om "Co-occuring" gener, mens "Mutually exclusive" gener har en log2 Odds Ratio, der er mindre end nul.
- Når du har undersøgt resultaterne på fanen Mutual Exclusivity, skal du bruge dem i punkt G.

G. Ud fra din statistiske test skal du formulere en konklusion på en måde, som giver mening i en bioteknologisk sammenhæng. Hvad viste den statistiske analyse:
   - For hvilke genkombinationer er gensidig udelukkelse (*mutual exclusivity*) statistisk signifikant?
   - For hvilke genkombinationer er det statistisk signifikant, at de forekommer sammen?

H. Som en opsamlende konklusion: Er det nok, at ét af de tre gener *CDK4*, *RB1* eller *CDKN2A* er blevet ændret, for at ødelægge styringen af cellens vækst og deling via RB-signaleringskaskaden? Hvorfor/hvorfor ikke? Hvilken betydning har det for kræft?

I. Ekstra:
   - Er *RB1* et tumorsuppressorgen eller er det et proto-onkogen? Begrund svaret ved hjælp af denne vejlednings beskrivelse af *RB1*, samt definitionen af de to typer af gener i artiklen om kræftbiologi.
   - Er *CDK4* eller *CDKN2A* tumorsuppressorgener eller proto-onkogener?

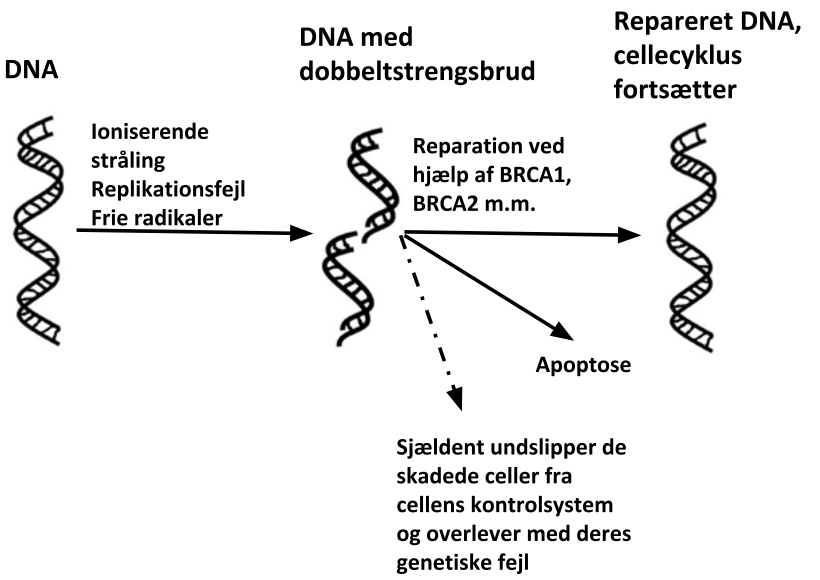
Biotekøvelse 3. Æggestokkræft: Mutationer i *BRCA1* og *BRCA2* og deres betydning for overlevelse

Formål

- at undersøge, om overlevelsestiden er kortere eller længere hos æggestokkræftpatienter med mutationer i generne *BRCA1* og *BRCA2* end hos patienter med andre mutationer.
- at diskutere mulige årsager til eventuelle forskelle i overlevelse.

Baggrund

Proteinerne BRCA1[1] og BRCA2[2] er i deres normale, velfungerende form med til at reparere dobbeltstrengs-brud i DNA (figur 1). Generne for de to proteiner hedder hver især det samme som proteinerne, nemlig henholdsvis *BRCA1* og *BRCA2*. Når dobbeltstrengsbrud og andre DNA-skader ikke repareres korrekt, fører det til ophobning af mutationer, og når disse mutationer findes i gener, der styrer cellens vækst og deling, medfører det øget risiko for kræft. Gener der koder for proteiner som bl.a. *BRCA1* og *BRCA2*, der bidrager til reparation af DNA-skader, kaldes tumorsuppressor gener (TSG). Navnet skyldes, at de modvirker kræft, ved at kode for proteiner, der bidrager reparation af skader og kontrol af cellens vækst og deling.



Figur 1. Når der opstår en DNA-skade som f.eks. dobbeltstrengsbrud, rekrutteres nogle af cellens DNA-reparationssystemer, herunder proteinerne BRCA1 og BRCA2. Hvis der ikke sker reparation, vil det i de fleste tilfælde føre til programmeret celledød (apoptose), mens skadede celler i nogle få tilfælde kan undslippe kontrolsystemet og overleve med skaden.

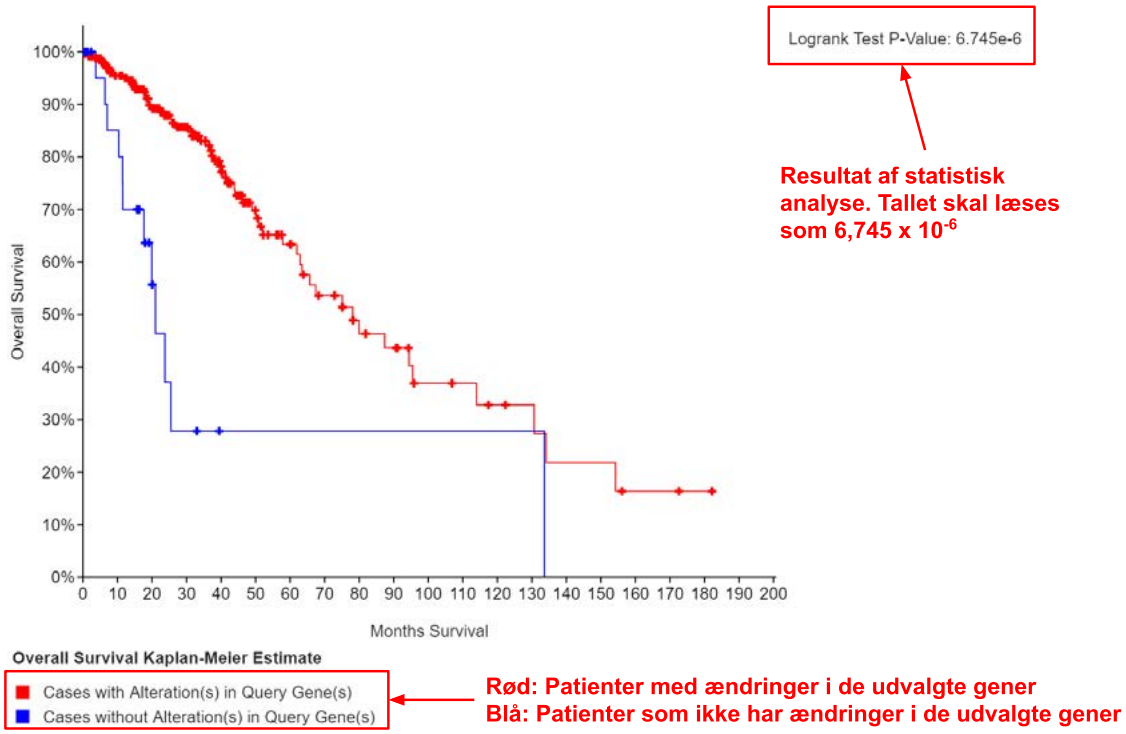
[1] https://www.genecards.org/cgi-bin/carddisp.pl?gene=BRCA1

[2] https://www.genecards.org/cgi-bin/carddisp.pl?gene=BRCA2&keywords=brca2

Nogle kræftformer som f.eks. brystkræft og æggestokkræft (ovariekræft) har en større arvelig disponering end andre kræftformer[3]. Det vil sige, at man kan arve et gen fra en eller begge af sine forældre, som øger risikoen for at få kræft. Muterede udgaver af *BRCA1*- og *BRCA2*-generne er sådanne gener. Det er vigtigt at lægge mærke til, at det netop er en *forøget risiko* for kræft, der er tale om. Som regel kræves der nemlig ændringer i flere gener, før man får kræft. Kun cirka 5% af kræfttilfældene i Danmark menes at være koblet til arvelige faktorer. Men da mutationer i generne *BRCA1* og *BRCA2* giver en stærkt forøget risiko for kræft i bl.a. bryst og æggestokke tilbydes en genetisk test for disse muterede gener samt ekstra undersøgelser til familiemedlemmer til patienter, der har en arvelig kræftdisposition.

Mutationer i *BRCA1* og *BRCA2* kan - udover at være nedarvet fra forældrene - også opstå i almindelige celler i løbet af livet, og så kaldes de *somatiske mutationer.* I modsætning hertil kaldes mutationer som stammer fra forældrene med et engelsk udtryk *germline mutationer*, hvilket henviser til, at det er muterede gener, der gives videre via forældrenes kønsceller (*germ cells* på engelsk), og dermed er de tilstede i alle afkommets celler lige fra befrugtningen.

Patienters overlevelseschance i forbindelse med kræft kan i nogle tilfælde relateres til specifikke genetiske ændringer, der findes som somatiske mutationer i patienternes tumorer eller som har været til stede som *germline* mutationer. Overlevelsen i forhold til specifikke mutationer undersøges nærmere i denne øvelse med data fra patienter med æggestokkræft. I øvelsen skelnes ikke mellem somatiske mutationer og germline mutationer. I figur 2 ses et eksempel på overlevelseskurver fra et studie med en anden kræftform end i denne øvelse.



Figur 2. Eksempel på overlevelseskurver fra et andet studie, end det, som der arbejdes med i denne øvelse. figuren stammer fra cBioPortal[4] med data fra studiet TCGA Provisional, om Lower-Grade Glioma. Kurver viser procentvis overlevelse af patienter (med en form for hjernekræft) siden første gang de blev diagnosticeret med kræft. Den røde kurve viser overlevelse af patienter med mutationer i de udvalgte gener *IDH1, IDH2* og *EGFR*. Den blå kurve viser overlevelse af patienter uden forandringer i netop disse gener.

[3] Dette afsnit er baseret på Kræftens Bekæmpelses hjemmeside www.cancer.dk, opslag om årsager til arvelighed, set 14. oktober 2019: https://www.cancer.dk/hjaelp-viden/fakta-om-kraeft/aarsager-til-kraeft/aarsager-arvelighed/

[4] Cerami, E. et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery, (2) (5) 401-404: https://cancerdiscovery.aacrjournals.org/content/2/5/401 samt Gao, J. et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal, Science Signaling pl1: https://stke.sciencemag.org/content/6/269/pl1

Find data

1. Gå til http://www.cbioportal.org/
2. I feltet "Search" indtastes det ønskede organ, i dette tilfælde "ovary".
3. Blandt søgeresultaterne vælges det ønskede studie: Sæt hak ud for studiet "Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)".
4. Vælg "Query By Gene"
5. Find "Enter genes"-boksen. I underboksen i midten, hvor der står "Enter HUGO gene symbols, Gene Aliases, or OQL" skal du taste gennavnene *BRCA1* og *BRCA2* (med almindelig skrift, ikke skrå skrift). Tjek at der under generne står med grøn baggrund "All gene symbols are valid". Hvis ikke, skal stavning af generne rettes.
6. Derefter trykkes på "Submit Query" og man venter.
7. Når søgningen er gennemført, vises automatisk fanebladet "OncoPrint". Skift til fanebladet "Comparison/Survival" og på den nye side skiftes til fanebladet "Survival", hvor overlevelseskurverne findes.
8. Download figuren og besvar arbejdsopgaverne:

Arbejdopgaver

A. Analyser overlevelseskurven som du fandt: Ser det ud til, at der er den samme overlevelse for patienter med æggestokkræft og *BRCA1*- eller *BRCA2*-mutationer i forhold til patienter som ikke har netop disse mutationer? Eller er der en dårligere eller bedre overlevelse? Formuler din foreløbige analyse som en hypotese, der skal testes.

B. Det du skrev i punkt A er en hypotese, som er formuleret med bioteknologisprog, og som skal testes med statistiske værktøjer. Selv om du måske har en klar fornemmelse af svaret ud fra figuren, skal du teste din hypotese med brug af data fra det studie, der ligger bag figuren. Hermed kan du afgøre, om der er *videnskabelig evidens*. Det kunne for eksempel være, at der var så stor variation inden for de to grupper, at det lige så godt kunne være tilfældigt, at det ser ud som om, der er forskelle.

Næste trin: statistisk analyse. Har klassen matematik med i forløbet?

Ja. Undersøg med statistiske metoder i matematiktimerne (Matematikøvelse 3) hvilken statistisk evidens der er for sammenhængen mellem overlevelse af patienter med *BRCA1*- eller *BRCA2*-mutationer i deres tumor, i forhold til patienter, der ikke har disse mutationer. Du kommer til at omdanne din bioteknologiske hypotese til en nulhypotese, der kan testes med matematiske værktøjer.

Når du er færdig med Matematikøvelse 3, skal du gå videre til næste punkt.

Nej. Så skal du ikke selv udføre den statistiske analyse, men bruge den, som cBioPortal laver: Der testes en nulhypotese om, at der ikke er forskel i overlevelsen i de to grupper.

- Hvis log-rank test *p*-værdien, som står på overlevelseskurverne, som du har fundet, er mindre end 0,05, forkastes nulhypotesen. Det vil sige, at de forskelle man kan se ikke skyldes tilfældigheder, og man kan konkludere, at der er forskelle.
- Hvilken retning en evt. forskel går kan ofte ses af figuren.
- Du skal altså nu bruge figuren og dens log-rank test *p*-værdi til at besvare næste punkt.

C. Hvordan lød nulhypotesen? Hvad blev resultatet af testen af nulhypotesen? Hvordan kan konklusionen på testen formuleres på bioteknologisprog?

D. Hvad kunne være årsagen til det, der blev fundet i punkt A og som blev understøttet af den statistiske test? (Overvej mulige årsager. Der er ikke et endeligt svar. En af ekstraopgaverne længere nede i vejledningen går ud på at undersøge, hvad nogle forskere har foreslået som årsagsforklaring).

E. Lav en samlet konklusion, som er en besvarelse af punkterne under "Formål" i starten af vejledningen. Husk at inddrage de statistiske resultater grundigt.

På næste side findes en række ekstraopgaver.

Ekstraopgaver

Opgaverne kan eventuelt fordeles på forskellige grupper, som laver en fremlæggelse i matrixgrupper.

Ekstra 1: Undersøg, om der findes den samme relation mellem *BRCA1*- og *BRCA2*- mutationer og overlevelse ved en anden kræftform, hvor *BRCA*-generne er involveret, nemlig brystkræft.

Brug samme fremgangsmåde som før (punkt 1-8) til at finde data, med følgende ændringer: Punkt 2: "breast", punkt 3 "Breast Invasive Carcinoma (TCGA, Firehose Legacy)".

Diskuter, hvad kurverne viser og sammenlign med resultatet for æggestokkræft. NB! Det er ikke sikkert, at der kan findes en årsagsforklaring til eventuelle forskelle eller ligheder. Det diskuteres i skrivende stund stadig i forskningsverdenen.

Ekstra 2: Undersøg mere om æggestokkræft og tests for *BRCA1* og *BRCA2*, samt om kræft og arvelighed generelt: Gå ind på Kræftens Bekæmpelses hjemmeside. Der er angivet links, men hvis de ikke skulle virke, kan man prøve at søge oplysningerne frem via hjemmesidens søgefunktion. Læs om

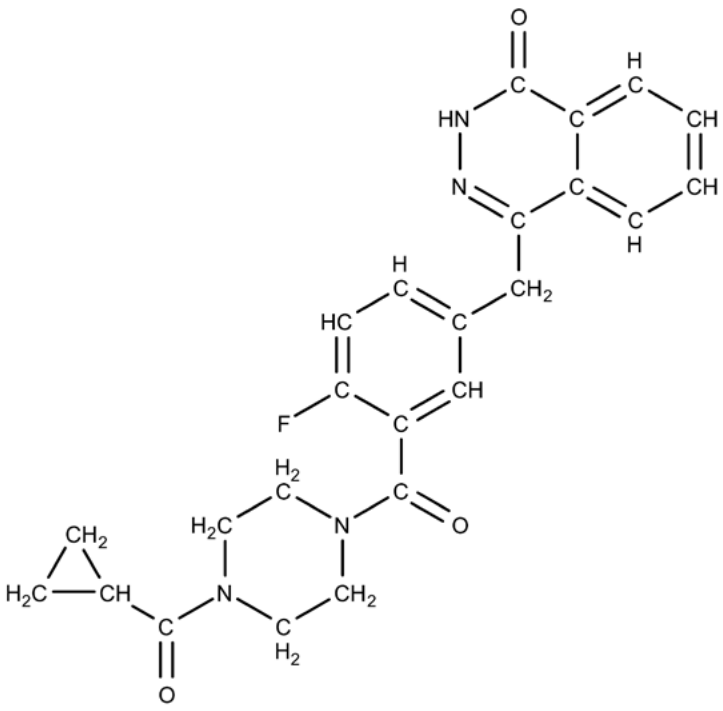
- æggestokkræft og tests for BRCA1 og BRCA2 m.m. (langt nede på siden)
- kræft og arvelighed generelt

Ekstra 3: Udvælg et enkelt emne inden for DNA-reparation i denne kilde, som dog ikke kommer ind på dobbeltstrengsbrud:
Lautrup, S. H., Stevnsner, T. (2015). Når kroppen reparerer DNA. Aktuel Naturvidenskab, 6, 28-33.

Ekstra 4: Undersøg mulige årsager til de fundne forskelle i overlevelse: Andre forskere end de, der står bag data i denne øvelse, har arbejdet med overlevelse af patienter med æggestokkræft. Undersøg deres arbejde i den nedennævnte kilde således: **Abstract** skal skimmes for at finde ud af, hvad forskerne selv har undersøgt og hvad de har konkluderet. **Indledningen til artiklen** skal skimmes indtil artiklens forfattere nævner mulige årsager til de observerede sammenhænge. Du skal altså IKKE læse hele artiklen. Brug denne kilde:
McLaughlin, J. R., et al. (2013). Long-term ovarian cancer survival associated with mutation in BRCA1 or BRCA2. Journal of the National Cancer Institute, 105(2), 141–148.

Ekstra 5: I figur 4 ses den kemiske struktur af aktivstoffet olaparib, som i nogle tilfælde anvendes til behandling af æggestokkræft.

- Opskriv sumformlen for molekylet på figur 4. Du må gerne anvende MarvinSketch eller lignende.
- Analyser molekylets evne til at trænge gennem cellemembranen.
- Olaparib er en såkaldt PARP-hæmmer. Undersøg hvordan PARP-hæmmere virker, for eksempel på "pro.medicin - Information til sundhedsfaglige"



Figur 4. Den kemiske struktur af olaparib, hentet fra MarvinSketch.

Biotekøvelse 4. Brystkræft: Genændringer og præcisionsmedicin

Formål

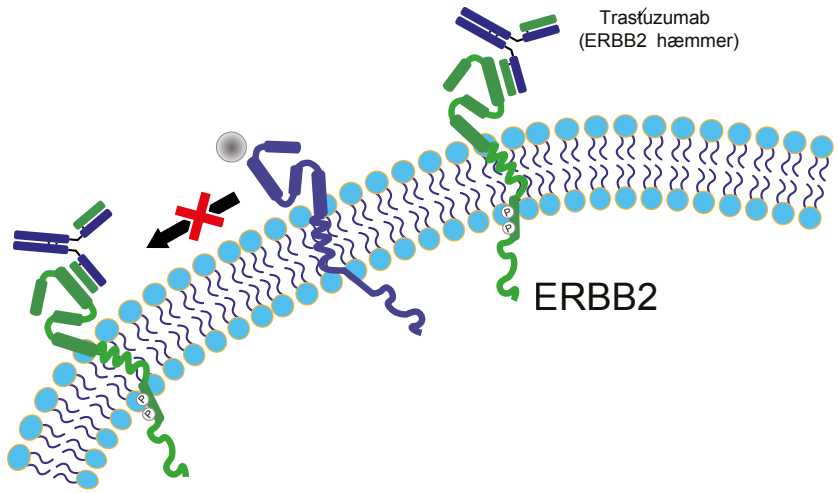
- at prøve at anvende en offentligt tilgængelig database, *cBioPortal*, som bl. a. bruges af sundhedspersonale (læger og molekylærbiologer), samt forskere inden for cancerområdet.
- at undersøge - i data fra studier af brystkræft - hvilke forskellige ændringer, der er i et bestemt gen, *ERBB2*.
- at undersøge hvordan præcisionsmedicin er relevant for patienter med brystkræft.

Baggrund

I Danmark diagnosticeres årligt ca. 4.650 nye tilfælde af brystkræft. I nogle tilfælde af brystkræft spiller genet *ERBB2* en rolle, og nogle patienter med ændringer i dette gen kan behandles med præcisionsmedicin. Genet *ERBB2* og den receptor, den koder for, samt genets og receptorens rolle i brystkræft er beskrevet i Biotekøvelse 1, samt i artiklerne om Kræftbiologi og Præcisionsmedicin.

Overekspression, altså en forøget transskription og translation af genet *ERBB2*, ses i tumorer hos nogle brystkræftpatienter. Overekspressionen er korreleret til amplifikation af det kodende gen. Med amplifikation menes, at der findes betydeligt flere kopier end normalt af genet, hvilket fører til et øget antal af receptorer i cellemembranen. Amplifikationer af genet *ERBB2* kan bruges til at forudsige respons på behandling, som indrettes derefter.

Et antistofpræparat med det virksomme stof trastuzumab (et monoklonalt antistof som indgår i medicin med handelsnavnet *Herceptin*) er specifikt designet til kun at binde sig til celler, som har øget ekspression af *ERBB2*. Når trastuzumab binder til ERBB2, formindskes tumorernes vækst ved at blokere ERBB2- proteinets funktion (figur 1) samt at tiltrække immunceller, der dræber kræftcellerne.



Figur 1. Virkning af antistofpræparatet Herceptin (aktivstof: Trastuzumab).
Herceptin virker ved at binde sig til receptoren ERBB2 (der også kaldes HER2). Derved kan vækstfaktor og den anden receptor ikke bindes til ERBB2 og signalkaskaden som vises i artiklen om kræftbiologi kan ikke aktiveres. Således stoppes overaktivering af celledelingen. Samtidig tiltrækker det bundne antistof, Herceptin, immunceller, så kræftcellerne dræbes.

Herceptin er et af de første medikamenter, som er blevet ordineret på baggrund af en molekylærbiologisk test[1]. Testen kan påvise, om en patient har overekspression af det 'mål', som medicinen er rettet mod, nemlig ERBB2-proteinet. Brystkræftpatienter, hvis tumorvæv *ikke* har overekspression af *ERBB2,* har ingen gavnlig effekt af *Herceptin*-behandling, og vil ved behandling kunne udvikle livstruende bivirkninger i form af f.eks. hjertesvigt.

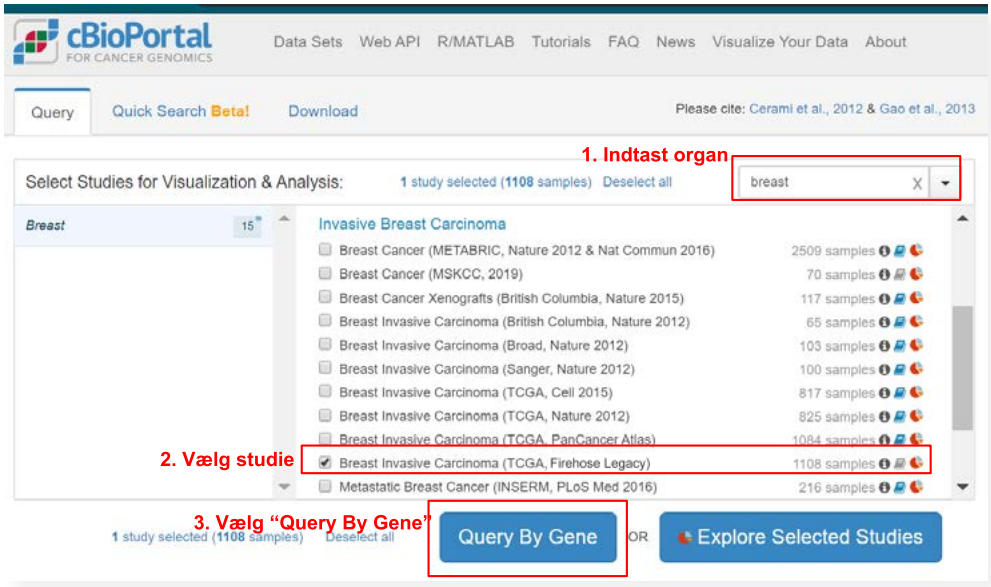
Du skal i denne opgave undersøge hvilke forskellige ændringer (amplifikationer og andre), der findes i genet *ERBB2* hos tumorer fra brystkræftpatienter, og hvor mange af disse patienter, der kan hjælpes med den nævnte medicin.

Fremgangsmåde og arbejdsspørgsmål.

Der er i alt 8 punkter i fremgangsmåden, og der er spørgsmål fra A-J.

Find datasættet

1. Gå til [cBioPortal's hjemmeside](). I feltet "Search" tastes organet som man vil arbejde med, her "breast" (figur 2).



Figur 2. Indtast organ i det felt på hjemmesiden, hvor der til at starte med står "Search". Sæt hak ved det ønskede studie og vælg "Query By Gene".

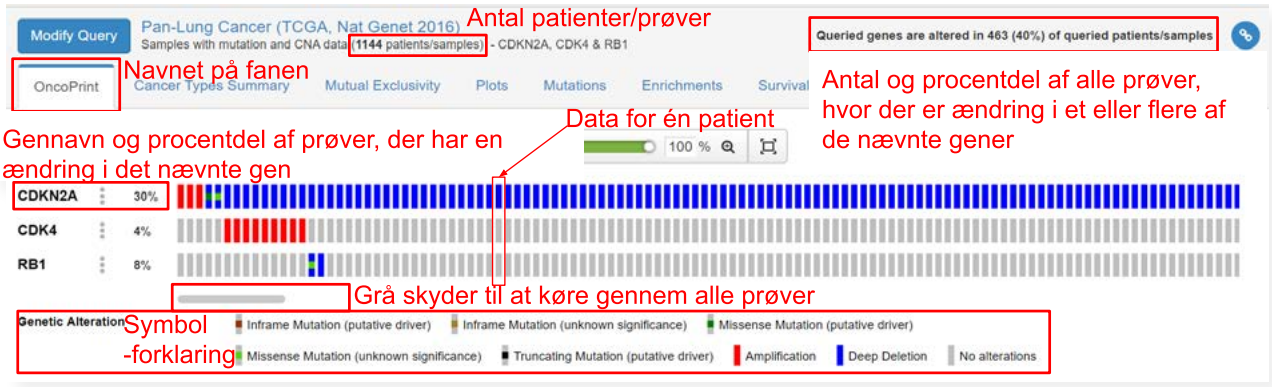
[1] En beskrivelse af de molekylærbiologiske tests for overekspression af *ERBB2*, som omfatter immunohistokemi og "fluorescens in situ hybridisering" (FISH) kan læses i [Nørager, M. T. et al. (2018) Sammenligning af automatiseret HER2 IQFISH med manuel HER2 FISH, Danske Bioanalytikere, nr 9, 22-25]()

2. Under "Invasive Breast Carcinoma" (som er en variant af brystkræft), sæt hak ved "Breast Invasive Carcinoma (TCGA, Firehose Legacy)" (figur 2). Nu har du valgt et bestemt videnskabeligt studie, som er udført med netop denne type af brystkræft.

3. Vælg "Query By Gene" (figur 2).

4. Find "Enter genes"-boksen. I underboksen i midten, hvor der står "Enter HUGO gene symbols, Gene Aliases, or OQL" skal du taste gennavnet *ERBB2* (med almindelig skrift, ikke skrå skrift). Tryk på "Submit Query".

5. Efter lidt ventetid kommer der et resultatvindue frem, som starter på fanen "OncoPrint". Det svarer til figur 3, som dog er fra et andet studie end det, som du skal arbejde med. I figur 3 og figurteksten dertil skal læses forklaringen til, hvordan figuren bruges. Besvar derefter arbejdsspørgsmålene, som kommer efter figuren.



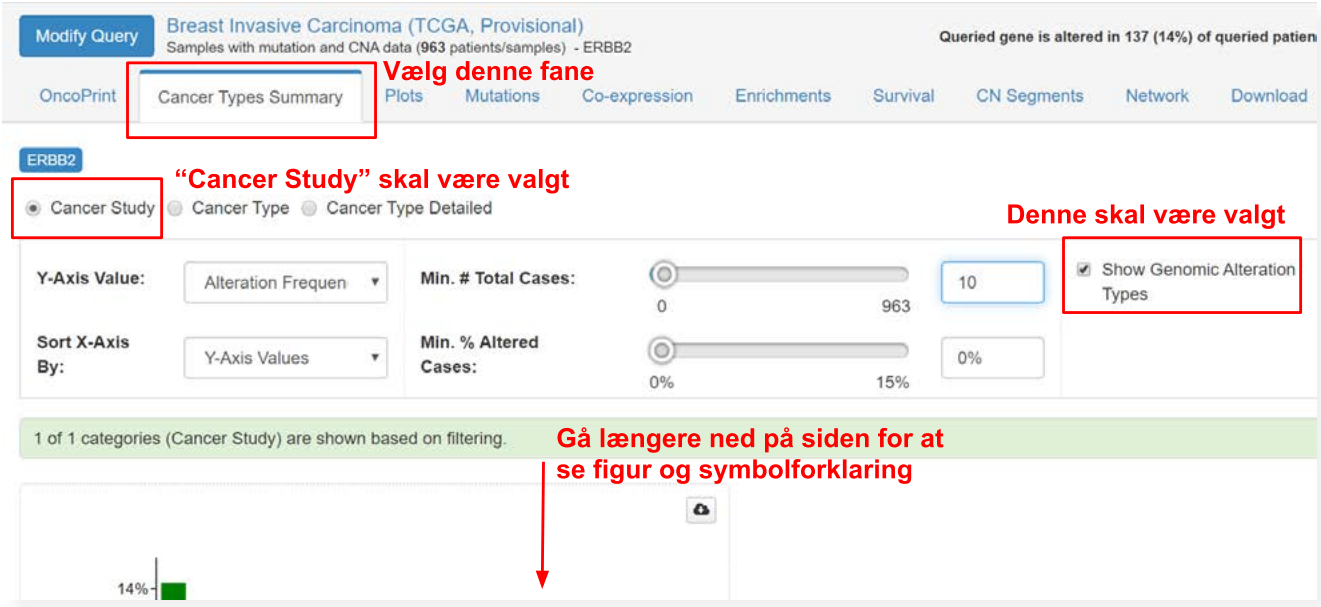
Figur 3. Resultat af søgning med de vigtigste områder markeret. Søgningen er på andre gener og med en anden type kræft end i denne øvelse. Billedet viser kun en lille del af undersøgelsens resultater, da der indgår mange patienter. Man skal sikre sig, at man har set information for alle undersøgte patienter: Hvis der er en grå skyder i billedet, så træk mod højre og venstre for at gennemgå alle patienternes data. Se markeringer på figuren for hvor man kan finde informationerne. Hver søjle ud for gennavnene repræsenterer en patient, hvis tumor er undersøgt for ændringer i generne. Procenttallet til højre for hvert gen angiver hvor stor en andel af alle de undersøgte patienter i studiet, der i deres tumor har en eller anden form for ændring af genet. Betydning af farvekoderne på hver patient står under rækken af patienter.

Undersøg hvilke ændringer, der er fundet i genet ***ERBB2*** hos bryst-kræftpatienter i et stort studie:

Nu har du set, at der findes forskellige ændringer af genet *ERBB2* i patienternes tumorer. Gå herefter videre med datasættet:

Undersøg hvor mange af patienterne, der kunne få gavn af midlet Herceptin.

6. Fortsæt med den samme undersøgelse som du lavede oven for. Gå til fanen "Cancer Types summary" (figur 4). Marker "Cancer study" og tjek, at Show Genomic Alteration Types er valgt.



Figur 4. Fanen Cancer Types Summary.

7. Hold cursoren hen over søjlen med farver, som er nederst i det skærmbillede, der kommer frem (kun toppen er vist på figur 4.)

8. Besvar spørgsmålene:

F. Hvilke brystkræftpatienter kan, ifølge indledningen til opgaven, få gavn af *Herceptin*?

G. Hvor mange procent af patienterne i dette studie ville kunne få gavn af behandling med *Herceptin*?

H. Hvordan stemmer procentandel fra det foregående punkt overens med antallet i denne tekst her? (Hvis linket ikke virker, kan oplysningerne søges frem på Kræftens Bekæmpelses hjemmeside cancer.dk på "behandling af HER2-positiv brystkræft ". *HER2* og *ERBB2* er det samme. Medicinen *Herceptin* har som aktivstof "trastuzumab")

I. Diskuter, hvorfor man kan bruge begrebet *præcisionsmedicin* i denne sammenhæng.

J. Diskuter fordele og ulemper ved brug af præcisionsmedicin til behandling af brystkræft.

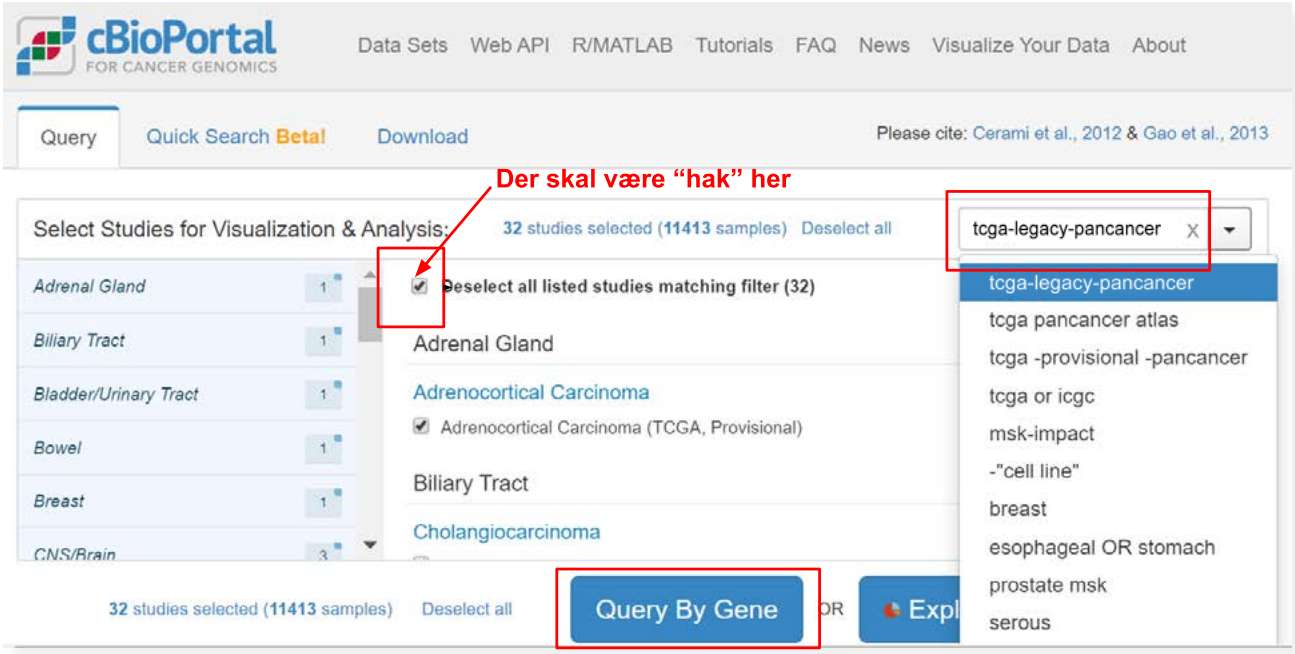
Ekstraopgaver

Der er to ekstraopgaver. Eleverne arbejder i grupper og efter endt gruppearbejde vælges ved lodtrækning de to grupper, der skal holde et kort oplæg om en af opgaverne med brug af figurer og fagbegreber. Til sidst står forslag til en opsamlende klassediskussion.

Ekstra 1. (Kan bruges, hvis klassen har lært om immunforsvaret). Undersøg mere om hvad lægemidlet *Herceptin* indeholder og hvordan det virker. Der skal bruges figurer og fagbegreber. Brug dine internetsøgningsskills.

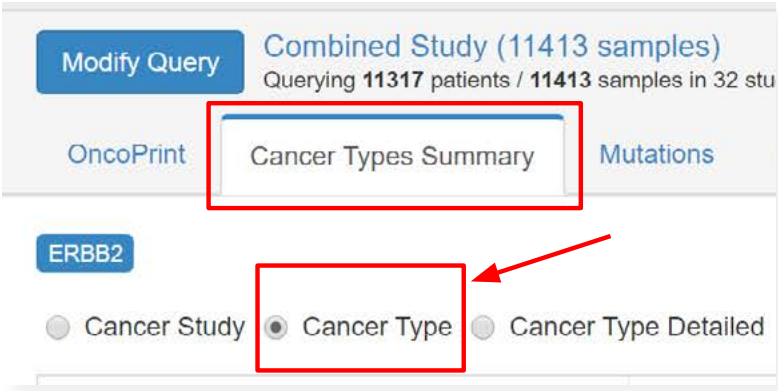
Ekstra 2. *Herceptin* virker, som set i ovenstående øvelser, godt mod brystkræft, som skyldes amplifikationer i *ERBB2*. Undersøg nu, om det samme gen er amplificeret i andre kræftformer, og undersøg, om den samme medicin på nuværende tidspunkt bruges rutinemæssigt som behandling ved andre kræftformer med samme amplifikation:

- Gå til cBioPortal forside.
- Under "Select Studies" vælges i dropdown-menuen (tryk på ned-pilen) i søgefeltet "tcga-legacy-pancancer" (figur 5) og derefter sættes der hak ved "Select all listed studies matching filter" (figur 5).



Figur 5

- Tryk på "Query By Gene".
- Find "Enter genes"-boksen. I underboksen i midten, hvor der står "Enter HUGO gene symbols, Gene Aliases, or OQL" skal du taste gennavnet ERBB2 (med almindelig skrift, ikke skrå skrift).
- Tryk "Submit Query"
- Vælg fanen "Cancer Types Summary" og vælg "Cancer Type" (figur 6).



Figur 6

- Se på figuren som kommer frem længere ned på siden, hvilke andre kræftformer, som har amplifikationer af *ERBB2* (den røde farve i søjlerne i figuren angiver andelen af patienter for hver kræftform i studiet, der har amplifikationer af *ERBB2*)
- Undersøg om *Herceptin* i dag bruges til behandling af andre kræftformer end brystkræft ved at søge på "Herceptin" eller "trastuzumab" på min.medicin.dk

Klassediskussion efter fremlæggelserne

Diskuter, for og imod, at *Herceptin* kan indgå i behandlingen af andre kræftformer end brystkræft

Diskuter, hvordan man kan forestille sig, at sundhedspersonale og molekylærbiologer kan arbejde med at undersøge, om *Herceptin* er en velegnet behandling til andre kræftformer: kom bl.a. ind på "eksperimentel behandling" og "evidensbaseret behandling" samt andre studier, der måtte gå forud for klinisk afprøvning.

Kræftbiologi, første del, gruppearbejde

Gruppearbejde på basis af første del af artiklen "Kræftbiologi"

4 opgaver, 8 grupper. Hver opgave laves af to grupper. Derefter fremlæggelse individuelt i matrixgrupper.

Del 1 af artiklen: Intro + Udvikling og celleprogrammering.

Find relevante figurer fra egen lærebog + evt. anden kilde til at fremlægge denne del af artiklen.

Del 2 af artiklen: Cellesignalering og vækstkontrol.

figuren "To virkemåder for onkogen aktivering af receptorer for cellesignaler" fra artiklen + evt. figurer fra egen lærebog bruges til at fremlægge denne del af artiklen.

Del 3 af artiklen: Vedligeholdelse af genomet, første to afsnit.

Find figurer til at fremlægge denne del af artiklen. Se først om der er noget at finde i egen lærebog, og søg derefter på internettet, brug kun kvalitetskilder.

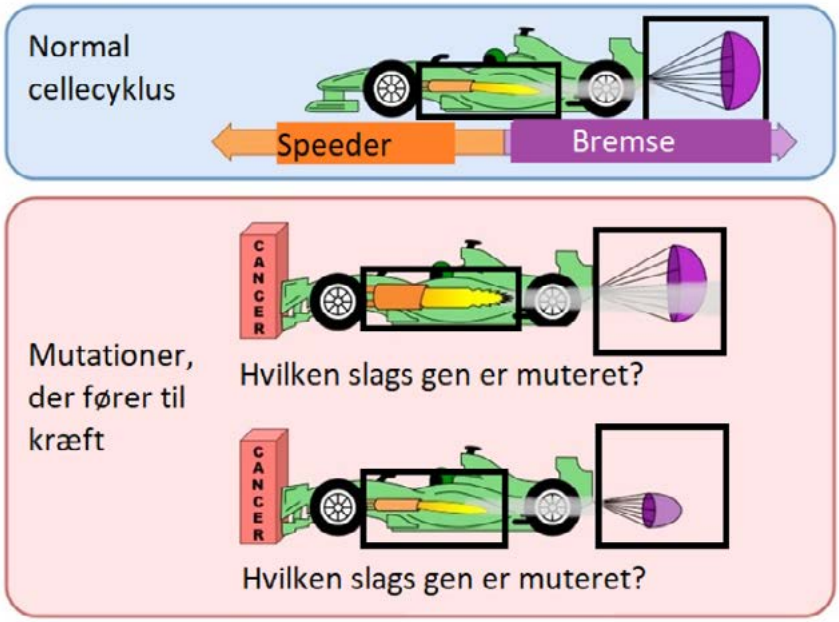
Del 4 af artiklen: Fra "Når en DNA-skade opstår" og resten af afsnittet.

Brug figuren "Telomerase". Eventuelt kan figuren fra en senere del af artiklen "Samspil mellem signalveje står for en kompleks kontrol af cellernes funktion" inddrages.

Kræftbiologi 2

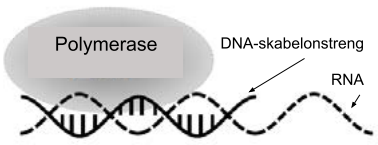
Arbejdsark til artiklen Kræftbiologi, fra og med overskriften "Kræft" og til slutningen af artiklen.

1. Hvad er et "drivergen"? Hvilke to overordnede typer findes der, og hvordan er de forskellige?
2. På figuren nedenunder er der illustreret normal styring af celledeling samt to forskellige resultater af mutationer i gener. Hvor er der sket mutationer i tumorsuppressorgener, og hvor er der sket mutationer i et proto-onkogen, som derved er blevet til et onkogen?



Figur 1. Styring af cellecyklus og mutationer i tumorsuppressorgener og proto-onkogener. Hvad er hvad? figuren er omtegnet efter https://ib.bioninja.com.au/_Media/oncogene_med.jpeg

3. Forklar dette nærmere: "Mens drivermutationer altid forekommer i drivergener, er det ikke alle mutationer, som forekommer i drivergener, som er drivermutationer?" (citat fra artiklen).
4. Hvilken proces i cellen illustrerer denne figur?



   a. Transskription
   b. DNA-replikation
   c. Translation

5. Hvordan vil en deletion af en central del af genet for et protein påvirke processen, som er vist i den lille figur i punkt 4, og hvordan vil det påvirke dannelsen af proteinet?
6. Hvorfor er det almindeligt at se i kræftceller, at *begge* udgaver af tumorsuppessorgener er ramt af mutationer?
7. Hvorfor er det almindeligt at se i kræftceller, at kun én udgave af et proto-onkogen er ramt af mutation, så det er blevet til et onkogen?
8. En kinase er et protein, der udfører en fosforylering af et protein. Hvad er en fosforylering og hvilken konsekvens har den som regel for proteinet, der bliver fosforyleret?
9. Fosforylering:
    a. Find et eksempel på fosforylering i figuren "Retinoblastom (RB1)-proteinet er hovedregulator af cellecyklus" fra artiklen.
    b. Forklar hvilken virkning fosforyleringen har i styring af cellecyklus.
10. Find RB i figuren "Retinoblastom (RB1)-proteinet er hovedregulator af cellecyklus" fra artiklen "Kræftbiologi". Brug figuren til at forklare, hvorfor cellecyklus bliver overstimuleret, hvis genet for proteinet RB er udsat for mutationer som gør, at RB-proteinet ikke længere virker som det skal. Diskuter to forskellige måder, som *man kunne tænke sig* at RB-proteinet kunne ændres på, så det ikke kan udføre sin normale funktion.

Ekstraopgaver

Ekstra-1: Undersøg forholdet mellem godartede (benigne) og ondartede (maligne) tumorer, og se, hvordan flere mutationer er nødvendige for udvikling af kræft: Læs på dette link under overskrifterne "Godartede knuder (benign tumor)" og "Ondartede knuder (malign tumor)". Læg mærke til figurerne. Overvej hvor man kan inddrage begreberne drivergener, tumorsuppressorgener, proto-onkogener og onkogener. *Hvis linket ikke virker, kan du gå til* Kræftens Bekæmpelses hjemmeside *og søge på "Hvad er kræft".*

Ekstra-2: (en svær opgave). Forklar mundtligt hele figuren "Retinoblastom (RB1)-proteinet er hovedregulator af cellecyklus" i artiklen med fagbegreber. Brug meget gerne også fagbegreber, som du kender, men som ikke står på figuren. Forklar også, hvorfor man i kræftceller som regel kun finder ét muteret gen inden for hver signalvej.

Ekstra-3: (en svær opgave). Forklar mundtligt hele figuren "Samspil mellem signalveje står for en kompleks kontrol af cellernes funktion" i artiklen med fagbegreber. Brug meget gerne også fagbegreber, som du kender, men som ikke står på figuren.

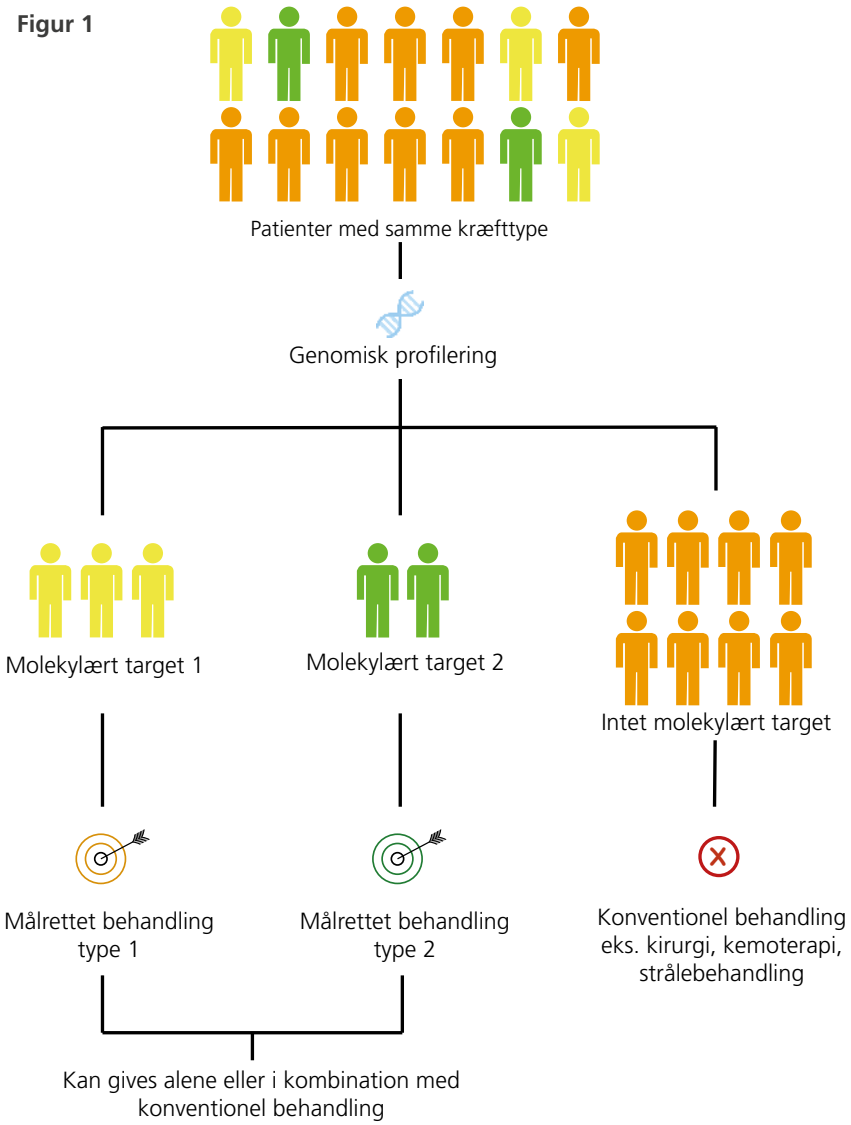
Præcisionsmedicin i kræftbehandling – Arbejdsark

Arbejdsarket besvares på basis af artiklen med samme titel.

Konventionel kræftbehandling forstås som brugen af operation, strålebehandling og kemoterapi. Kemoterapi er "Medicinsk behandling med stoffer, der standser cellers deling på forskellig måde (cellegifte). Denne type stoffer kaldes cytostatika. Rammer især de celler, der deler sig hurtigt, som kræftceller gør."[1]

1. Forklar mundtligt ved hjælp af nedenstående figur, hvordan præcisionsmedicin er forskellig fra konventionel kræftbehandling. Udelukker brugen af præcisionsmedicin den konventionelle behandling?

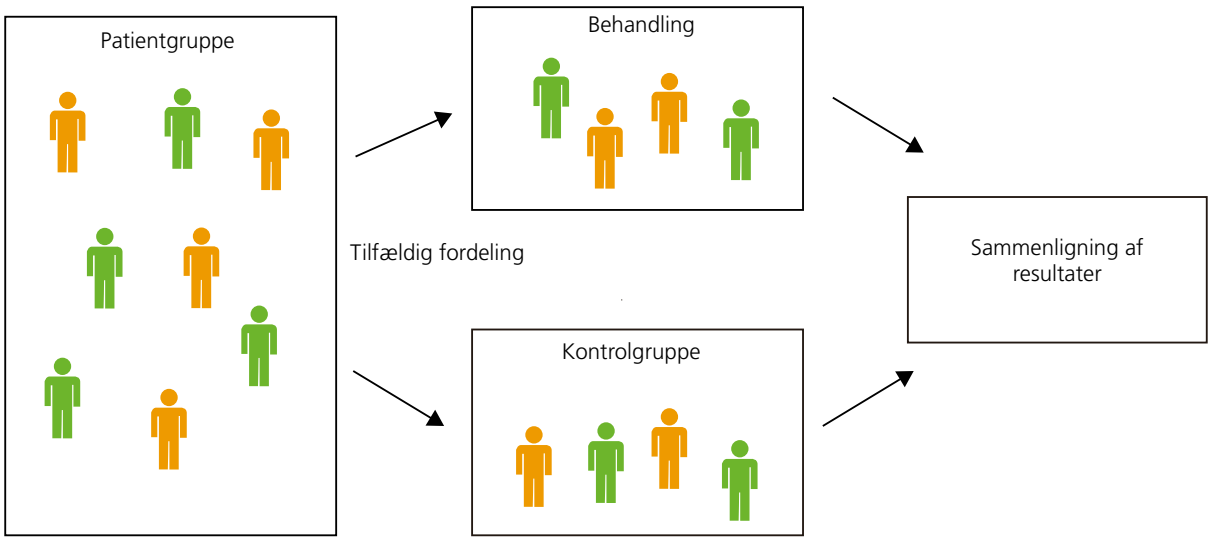
Figur 1



[1] https://www.cancer.dk/hjaelp-viden/fakta-om-kraeft/ordbog/K/, Kræftens Bekæmpelses ordbog, besøgt 2. juli 2019

2. Er dette udsagn sandt eller falsk?:
   "Præcisionsmedicin er det samme som at der skræddersyes *en helt unik behandling* til hver enkelt patient". Hvis det er sandt, så uddyb forklaringen. Hvis det er falsk, så giv en mere korrekt definition af præcisionsmedicin.

3. Randomiserede, kontrollerede forsøg:
   a. Forklar mundtligt ved hjælp af figur 2 hvad der ligger i begrebet "Randomiserede, kontrollerede forsøg". Det må gerne være mere uddybende end teksten i artiklen. Hvilke fordele og ulemper kan der være ved disse forsøg?
   b. Hvorfor er det ofte vanskeligt eller umuligt at lave store, kontrollerede undersøgelser med præcisionsmedicin?

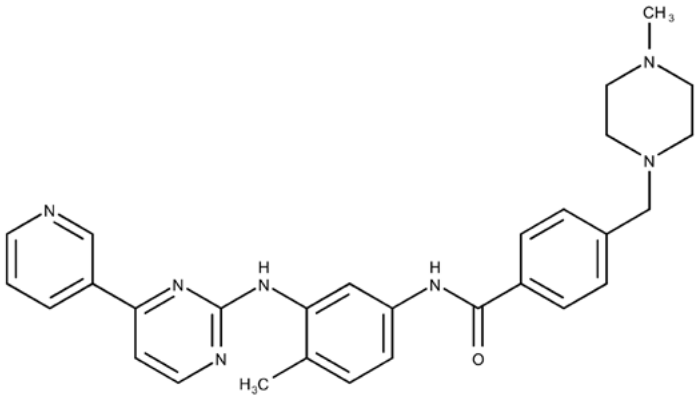


Figur 2. Randomiseret, kontrolleret forsøg.

4. Hvorfor er det typisk nemmere at lave medicin der rammer målproteiner, som onkogener koder for, end at lave medicin, der rammer celler, der har mutationer i tumorsuppressorgener?

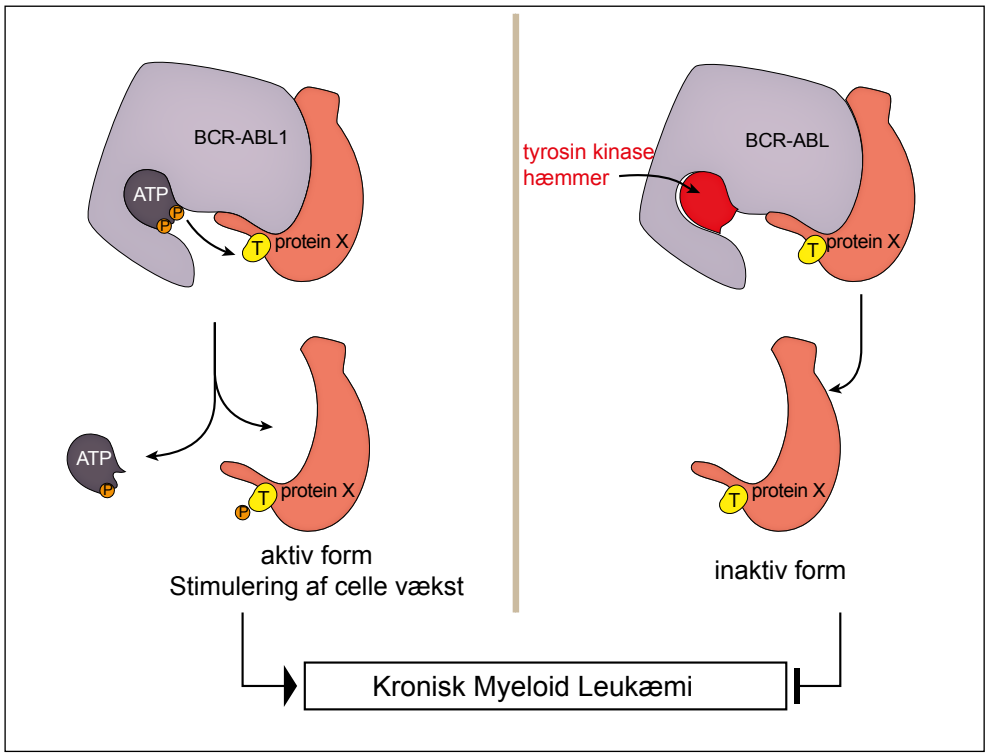
5. Forklar mundtligt NGS i forhold til den klassiske Sanger-sekventering. Inddrag figuren om sekventering fra artiklen.

6. Imatinib er en småmolekylær hæmmer, som anvendes mod CML, hvor patienten har Philiadelphia-kromosomet. figur 3 viser den kemiske struktur af lægemidlet samt en række kemiske egenskaber.
   a. Diskuter, om den kemiske struktur af imatinib tillader at stoffet transporteres gennem cellemembranen.



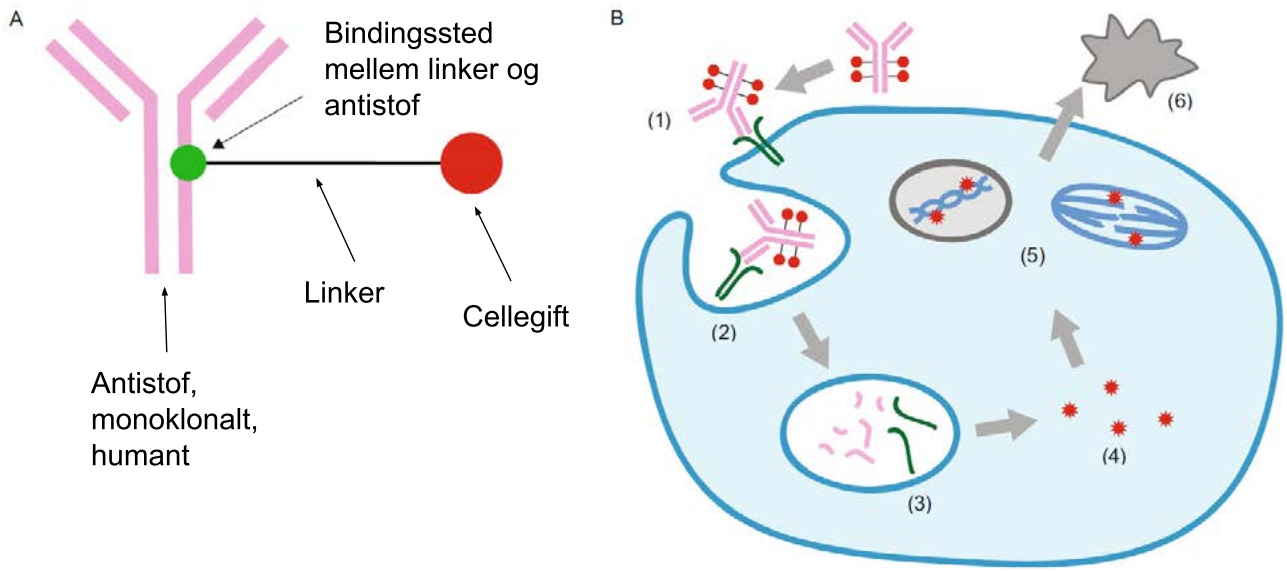
Figur 3. Den kemiske struktur af imatinib (MarvinSketch). Molekylformlen er $C_{29}H_{31}N_7O$ og molarmassen er 493,6 g/mol (MarvinSketch). Der er 7 hydrogenbindingsacceptorer, 2 hydrogenbindingsdonorer og en logP på 4,38.
(Kilde: https://www.drugbank.ca/salts/DBSALT000098).

7. Forklar mundtligt, hvordan et lægemiddel, der er en tyrosinkinase-hæmmer (f. eks. Imatinib), kan bruges til at hæmme kræftcellers vækst hos patienter med CML og Philadelphiakromosomet, idet figur 4 her fra arbejdsarket inddrages.



Figur 4

8. Forklar mundtligt: Hvilken strategi kan bruges til at få et lægemiddel frem til målet i cellen, hvis lægemidlet ikke selv kan bevæge sig gennem cellemembranen? Inddrag artiklens tekst og figur 5 her i arbejdsarket og forklar fagudtrykkene som bruges.



Figur 5. Antistofkonjugeret cellegift. Omtegnet efter Tsuchikama, K. & An, Z. Protein Cell (2018) 9: 33.
Oprindelig licens https://creativecommons.org/licenses/by/4.0/

9. Skriv en kort og præcis opsummering om præcisionsmedicin til kræftbehandling på 15-20 linjer. Målgruppe: En biotekelev i 3g på et andet gymnasium, som har arbejdet med kræft, men ikke med præcisionsmedicin. Følgende (fag)udtryk skal indgå – dog må to udelades efter eget valg:

Gruppering af patienter, eksperimentel behandling vs. evidensbaseret behandling, NGS, tumorens genetiske profil, tumorsuppressorgen, onkogen, signalvej, kontrol af cellecyklus, småmolekylær hæmmer, antistof.

Quizlet, beskrivelse og links

Hvad er Quizlet?

Dette materiale indeholder en række quizlets. En "Quizlet" er en slags onlinequiz, hvor begreber og forklaring af begreber trænes og evt. selvevalueres. Når først de matchende begreber og forklaringer er til rådighed som i dette materiale, er der mange forskellige muligheder for at træne og teste, både almindeligt terperi og mere spilagtige aktiviteter. Der gives automatisk feedback. Eleverne skal blot følge links i materialet, hvor de bliver bedt om at registrere sig som brugere, hvilket er gratis i skrivende stund. Læreren skal huske at vælge at registrere sig som lærer, hvilket også er gratis til det basale, selvom man nok ikke undgår at få reklamer for opgradering til betalingsversionen.

Hvad er Quizlet Live?

Quizlet Live er et sjovt og lærerigt online spil som kan bruges i timerne med fokus på samarbejde i grupper. Udgangspunktet er de sæt af begreber og definitioner, Quizlets, som blandt andet stilles til rådighed i dette materiale.

Eleverne arbejder sammen i automatisk dannede grupper om at kombinere 12 termer og definitioner. Det første hold, der kombinerer alle termer og definitioner korrekt, har vundet.

Som udgangspunkt skal man have et sæt med mindst 12 termer og definitioner. De fleste Quizletsæt i dette materiale har termer nok til Quizlet Live.

For at lade eleverne arbejde med Quizlet Live skal du registrere dig som lærer på Quizlet. Her er link til Quizlet hvor du kan oprette dig gratis - husk at vælge at du er lærer: https://quizlet.com/en-gb

Link til vejledning til Quizlet Live: https://quizlet.com/en-gb/help/2444125/how-to-use-quizlet-live

Quizlets til artiklen Kræftbiologi

Alle begreber til artiklen Kræftbiologi samlet (28 begreber)

https://quizlet.com/_6r2qvo

Begreber til artiklens første del, fra start til og med afsnittet "Vedligeholdelse af genomet" (15 begreber)

https://quizlet.com/_6r2r06

Begreber til artiklens sidste del, fra og med afsnittet "Kræft" (13 begreber)

https://quizlet.com/_6r2t9y

En lille, enkel quizlet hvor begreberne cellemembran, vækstfaktor, receptor, signaleringskaskade og celle-cyklus kobles til en figur. Denne er IKKE brugbar til Quizlet Live.

https://quizlet.com/_6rf9er

Quizlet til artiklen Præcisionsmedicin i kræftbehandling

Begreber til hele artiklen

https://quizlet.com/_6umcek

Fra mutation til præcisionsmedicin

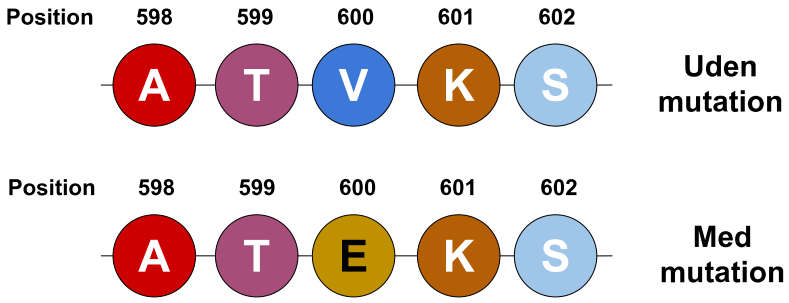
Typeord i opgaverne er ifølge Lærerens hæfte, Bioteknologi A STX, 2019

Opgaverne kan besvares som en skriftlig aflevering (90 minutter) eller som en screencastaflevering, hvor der peges på figurerne samtidig med besvarelsen af spørgsmålene. Screencasten må vare op til fem minutter.

BRAF V600E ved modermærkekræft

I nogle former for modermærkekræft findes ændringer i enzymet BRAF, som indgår i en signaleringskaskade, som stimulerer celledeling. Den normale BRAF katalyserer i den aktive form omdannelsen af det næste trin i signaleringskaskaden, MAP2K1, til en aktiv form.

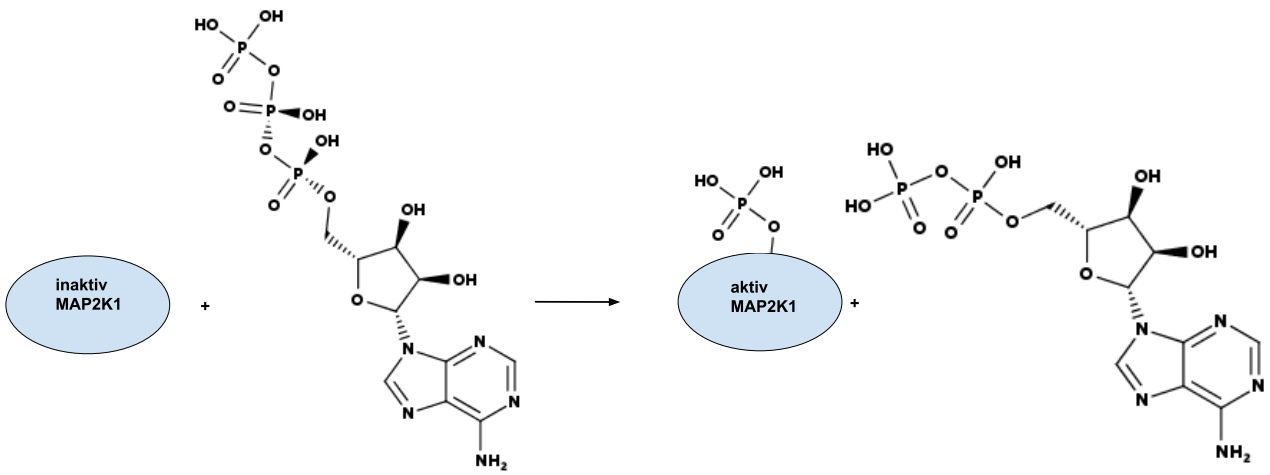
Den normale form af BRAF aktiveres selv ved, at et andet enzym tilføjer en fosfatgruppe til BRAF, og BRAF deaktiveres, når fosfatgruppen igen fjernes. Den ændrede form af enzymet, som kaldes BRAF V600E, er altid aktiv, og fosfatgruppen har altså ikke betydning for, om enzymet er aktivt eller ikke aktivt. Ændringen fra den normale til den kræftrelaterede form af BRAF skyldes, at en enkelt aminosyre i position 600, talt fra N-terminalen, er udskiftet (figur 1).



Figur 1. Udsnit af proteinet BRAF uden og med mutation i koden for aminosyre nummer 600. Aminosyrerne er angivet med deres et-bogstav-kode.

1. **Foreslå** hvilke(n) ændring(er) der kan være sket på det genetiske niveau, som har ført til ændringen af enzymet.

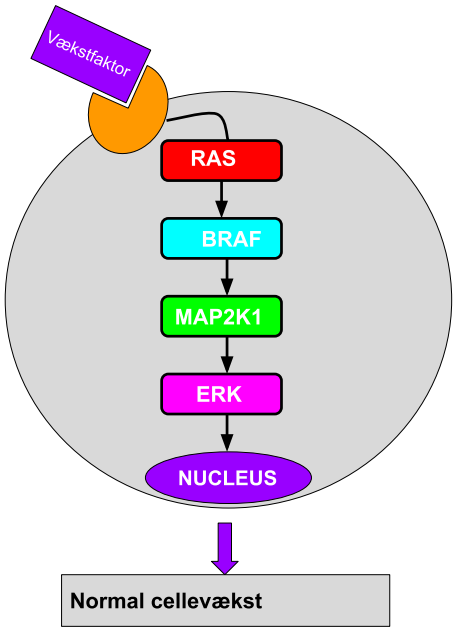
I figur 2 vises den reaktion, som det aktive BRAF katalyserer.



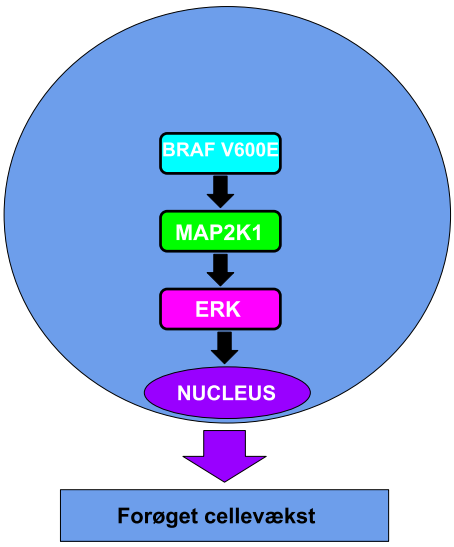
Figur 2. Reaktionen, som katalyseres af aktivt BRAF (ikke afstemt). Kemiske strukturer er fra MarvinSketch.

2. **Angiv** enzymtypen for det aktive BRAF-protein ud fra figur 2, **med en kort begrundelse**.

I figur 3a ses den normale funktion af signaleringskaskaden, som BRAF indgår i, mens funktionen med BRAF V600E fremgår af figur 3b.



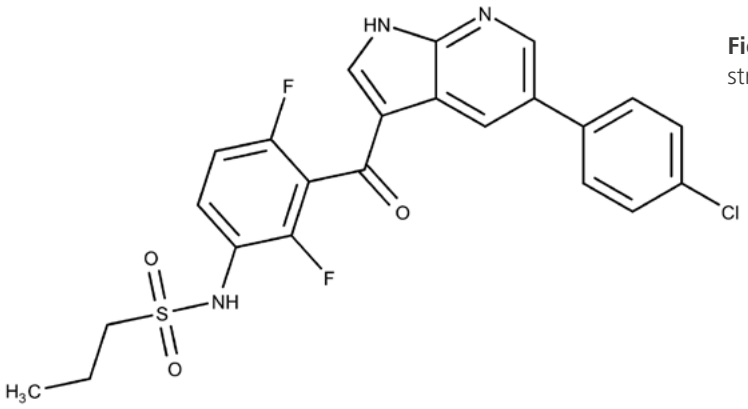
3a Signaleringskaskade med normal BRAF



3b Signaleringskaskade med BRAF V600E

3. **Forklar** hvordan ændringen af BRAF til BRAF V600E kan føre til ændringer i styringen af cellens vækst, idet figur 3a og b inddrages.

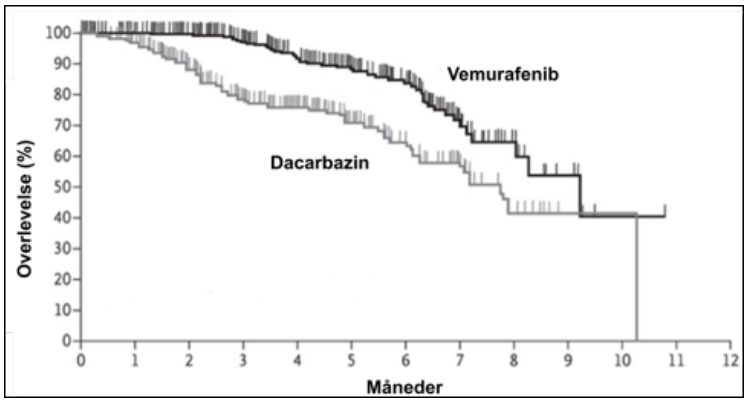
Der findes en præcisionsmedicin af typen småmolekylær hæmmer, som er blevet testet på modermærkekræftpatienter, hvis kræftvæv indeholder BRAF V600E. Medicinen kaldes vemurafenib (figur 4), og den skal trænge helt ind i cellen for at binde sig til BRAF V600E, og derved inaktivere enzymet.



Figur 4. Vemurafenibs kemiske struktur (fra MarvinSketch)

4. **Forklar** hvorfor vemurafenibs kemiske struktur tillader transport gennem cellemembranen. Inddrag figur 4.

I 2011 var præparatet dacarbazin, som bruges til kemoterapi, det eneste stof, der var godkendt i USA til behandling af modermærkekræft med metastaser. Vemurafenib har nu gennemgået en fase-3 klinisk afprøvning, hvor virkningen på overlevelsen blev sammenlignet med virkningen af dacarbazin (figur 5).



Figur 5. Kaplan-Meier overlevelseskurver for patienter, som er behandlet med henholdsvis Dacarbazin og Vemurafenib. De lodrette streger angiver, at data er censurerede (udtryk fra den statistiske behandling). Hver gruppe bestod af 336 patienter. Kilde: P. Champan et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. N Engl J Med 2011; 364:2507-2516

5. **Analyser** resultaterne i figur 5 og **diskuter** muligheden for anvendelse af Vemurafenib til behandling af modermærkekræft. Du skal gå ud fra, at forskellen mellem overlevelseskurverne i figuren er blevet analyseret med statistiske metoder, og at forskellen er signifikant.

Korrelationer og statistisk analyse

Anne Lyngholm, Bjarke Hansen og Claus Thorn Ekstrøm
| November 2019

Det skal vi lære

Når I har været igennem materialet skal I kunne svare på følgende:

- Forklare hvad Pearsons korrelationskoefficient er, og beskrive, hvad den måler.
- Forstå hvilke værdier, korrelationskoefficienten kan antage, og hvad de forskellige værdier betyder.
- Vurdere om antagelserne for at lave en korrelationsanalyse er overholdt.
- Forklare hvad konfidensintervallet for korrelationskoefficienten fortæller.
- Perspektivere til en lineær regression og forstå forskellen mellem de to metoder.

Introduktion

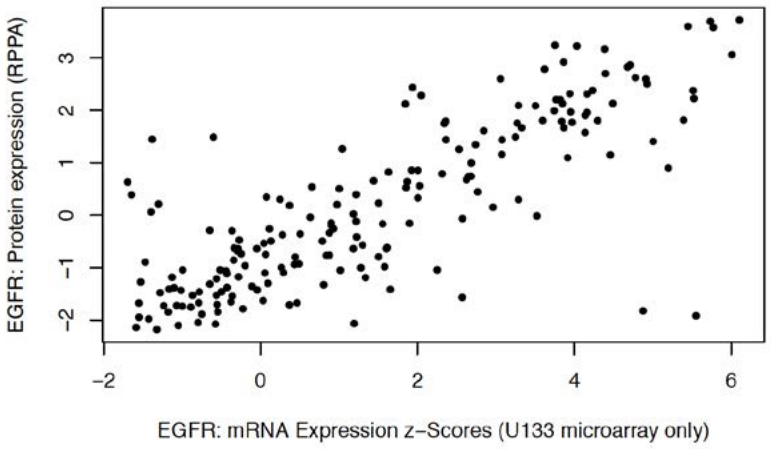
Vi skal her se på *Pearson korrelationen*, der er en statistisk metode til at bestemme, hvor stærk den lineære sammenhæng mellem to kontinuerte variable er. Metoden adskiller sig fra almindelig lineær regression ved, at de to variable optræder ligeværdigt: i lineær regression tænker man, at ændringer af variablen på x-aksen giver en sammenhæng med variablen på y-aksen. Ved korrelationsanalyse kan de to variable byttes rundt, og det er ikke oplagt, hvilken variabel der hører til x-aksen, og hvilken der hører til y-aksen. Korrelationen beskriver altså en symmetrisk lineær sammenhæng mellem de to variable. Vi skal se, hvordan man udregner korrelationen, hvordan man bruger resultatet til at vurdere den lineære sammenhæng, og hvilke antagelser der skal være opfyldt for at vi må udregne korrelationen.

Eksempel: sammenhængen mellem proteinekspression og mRNA ekspression for genet *EGFR*.

Vi ønsker at undersøge, om der er en lineær sammenhæng mellem proteinekspressionsniveauet og mRNA ekspressionsniveauet for genet *EGFR*. For at undersøge dette har vi indsamlet data fra en række personer, og har tegnet de standardiserede ekspressionsniveauer op mod hinanden. Den første del af data er gengivet i figur 1, hvor hver række har et id samt en måling af x (mRNA ekspression) og y (proteinekspressionsniveau). Sammenhængen mellem de to ekspressionsniveauer er vist i figur 2.

	A	B	C
1	Id	mRNA	RPPA
2	TCGA-02-0003-01	-0.5461	-1.844851
3	TCGA-02-0004-01	1.653	-1.414229
4	TCGA-02-0011-01	-1.4719	-0.894583
5	TCGA-02-0014-01	-0.7486	-1.888635
6	TCGA-02-0068-01	5.5499	-1.918899
7	TCGA-02-0069-01	-0.3435	-0.855257
8	TCGA-02-0116-01	0.2899	-1.09331
9	TCGA-02-2470-01	-0.3776	-1.651596

Figur 1. Udsnit af patientdata, der indeholder Id, mRNA ekspression og proteinekspressionsniveau (beregnet via RPPA)



Figur 2. *xy*-plot over sammenhængen mellem proteinekspression og mRNA ekspressionen for *EGFR*.

Figur 2 viser en positiv lineær sammenhæng mellem proteinniveauet og mRNA ekspressionen. Havde vi byttet rundt på de to variable (og dermed på de to akser), havde vi set helt samme mønster: et højt proteinniveau er ofte associeret med en tilhørende høj mRNA ekspression og omvendt. Da de to variable visuelt lader til at have en lineær sammenhæng, giver det mening at udregne korrelationen til at beskrive graden af den lineære sammenhæng.

Korrelationen

Korrelationen beskriver styrken af den lineære sammenhængen mellem to variable, og korrelationen måler, hvor tæt punkterne ligger omkring en ret linje. Derfor bør man altid starte med at tegne de to variable op mod hinanden i et *xy*-plot for at se, om sammenhængen er nogenlunde lineær. På et *xy*-plot kan man desuden med det samme se, om der er nogle punkter, der skiller sig markant ud fra de andre. Disse punkter kaldes for outliers. Da korrelationen kun kan måle lineære sammenhænge, kan *xy*-plottet opfattes som starten på en vurdering af metodens antagelser.

Hver måling optræder i par, hvilket også fremgår af eksemplet i figur 1, hvor hver række har et id samt en måling af *x* (mRNA ekspression) og *y* (proteinekspressionsniveau). For at kunne holde styr på, hvilke observationer, der hører sammen, skriver vi det *i*'te observationspar som: (y_i,x_i), hvor i = 1,..., *n*.
Det betyder, at y_1 og x_1 er det første observationspar og hører til den samme person (Id TCGA-02-00003-01 på figur 1), y_2 og x_2 er det andet observationspar og hører til en ny person osv. indtil vi til sidst har y_n og x_n, som er det *n*'te observationspar hørende til den sidste person.

Pearson korrelationen betegnes r_P og udregnes ved nedenstående formel:

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Bemærk, at i formlen indgår $\overline{x}$ og $\overline{y}$,der er gennemsnittet af henholdsvis *x*'erne og *y*'erne. Der gælder altså, at:

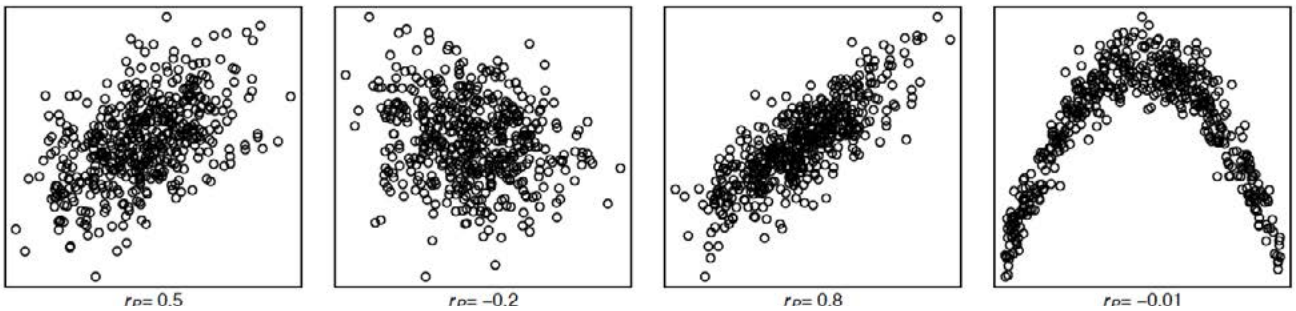
$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n},$$

hvor vi har brugt, at $\sum_{i=1}^{n} x_i$ står for summen af *x*'erne fra den første observation *i*=1 op til den sidste observation *i*=*n* og kan skrives som $x_1 + x_2 + x_3 + \cdots + x_n$.

Korrelationen kan have værdier mellem -1 og 1. En korrelation på -1 betyder, at der er en perfekt negativ lineær sammenhæng mellem to variable, en korrelation på 0 betyder, at der ikke er nogen lineær sammenhæng mellem de to variable, mens en korrelation på 1 betyder en perfekt positiv lineær sammenhæng mellem de to variable.

En korrelation tæt på 0 behøver ikke at betyde, at der ikke er en sammenhæng mellem de to variable, det betyder blot, at sammenhængen ikke er en særlig kraftig *lineær* sammenhæng.

Figur 3 viser fire eksempler på *xy*-plot med tilhørende Pearson korrelation beregnet ud fra overstående formel.



Figur 3.

Fire *xy*-plot som viser eksempler på forhold mellem to variable med hver deres estimerede Pearson korrelationsparameter r_p .

På de første tre plots af figur 3 ses, at jo tættere punkterne er på en ret linje, jo tættere er korrelationsparameteren på 1 (eller -1). Hældningerne i det første og det tredje plot er nogenlunde ens, men korrelationerne er forskellige, for korrelationen afhænger ikke af hældningen, men udelukkende af, hvor tæt punkterne ligger omkring en ret linje.

Det fjerde plot i figur 3 viser, at korrelationsparameteren ikke fortæller, om der er sammenhæng mellem de to variable, men kun fortæller, om der er en *lineær* sammenhæng. Der er en meget klar sammenhæng mellem de to variable på det sidste plot, den er bare ikke lineær (det giver faktisk slet ikke mening at beregne en korrelationsparameter for det sidste plot, efter at man har tegnet data, da man med det blotte øje kan konkludere, at der netop ikke er en lineær sammenhæng).

Eksempel: protein- og mRNA ekspression for *EGFR*

Figur 2 viste en positiv lineær sammenhæng og udregningen af r_p giver da også en Pearson korrelation på 0.80.

Jo større den numeriske værdi af korrelationskoefficienten er, jo tættere ligger punkterne omkring en linje, og vi skal senere se på, hvordan vi statistisk kan vurdere, om vi synes, at korrelationskoefficient er så langt fra 0, at vi kan konkludere, at der er en eller anden form for sammenhæng mellem de to variable.

Hvad skal være opfyldt, før man må udregne korrelationen?

Det er altid muligt at indsætte observationer i formlen for korrelationen og udregne resultatet, men der er nogle antagelser, der skal være opfyldt for, at det overhovedet giver mening at udregne korrelationen. Hvis antagelserne ikke er opfyldt, så vil resultaterne muligvis være unøjagtige, misvisende eller direkte forkerte.

- **Linearitet.** Det første skridt er at checke, om det giver mening at beskrive sammenhængen med en ret linje. Dette gøres ved at tegne et *xy*-plot, og ud fra plottet vurdere, om der er en tilnærmelsesvis lineær sammenhæng mellem de to variable.
- **Uafhængighed.** Observationsparrene (vores data) skal være uafhængige. At observationerne er uafhængige betyder, at et sæt målinger ikke giver os yderligere oplysninger om, hvad vi forventer at se ved næste par af målinger.

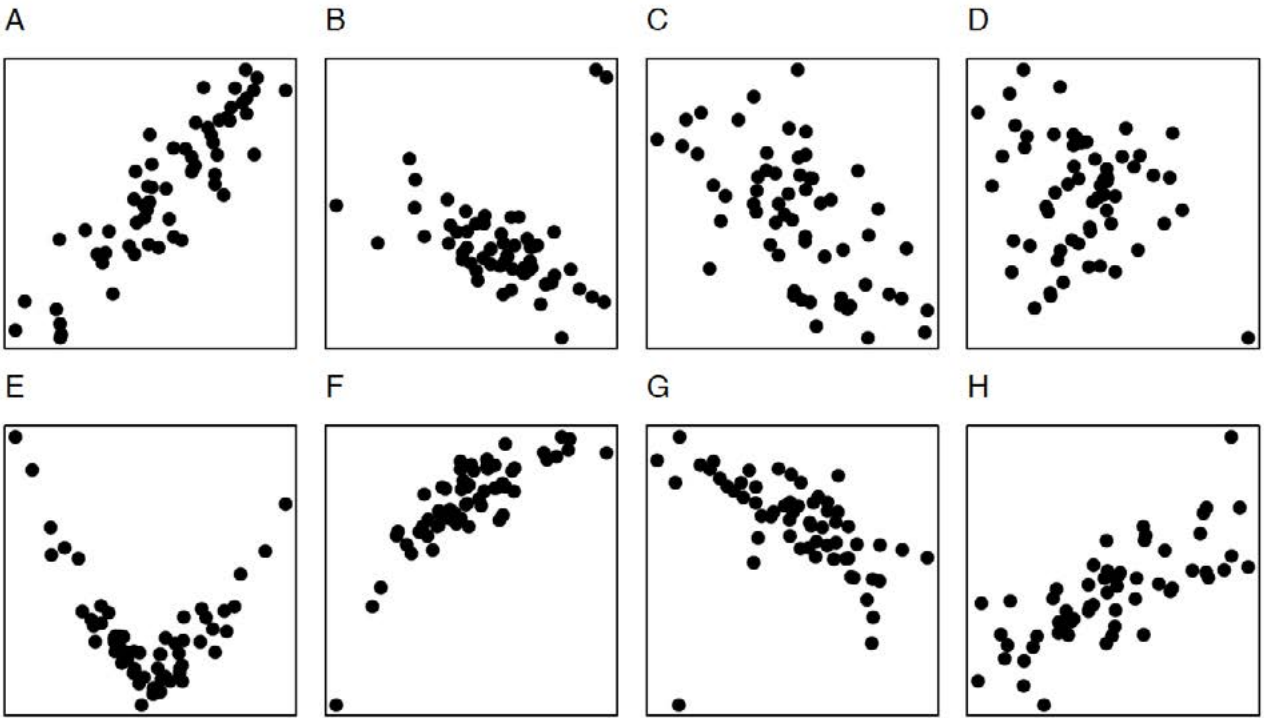
  Der er som regel kun en måde at undersøge uafhængighed på, og det er ved at overveje den metode, der er brugt til at indsamle data. Hvis man har taget en tilfældig og repræsentativ stikprøve fra en stor population, så vil uafhængighedsantagelsen (typisk) være opfyldt.

  Problemer med uafhængighed kan opstå, hvis man for eksempel måler hver person flere gange eller måler personer fra samme familie. Konsekvensen er, at man får forkerte usikkerheder på ens estimater, og at man tror, at estimaterne er mere præcise end de i virkeligheden er.
- **Ingen ekstreme outliers.** Outliers kan ses med det "blotte øje", da de stikker ud fra mængden på et *xy*-plot. Fordi en måling er anderledes end de øvrige målinger betyder det ikke, at den nødvendigvis er forkert.

  Hvis målingen er en målefejl eller er biologisk urealistisk værdi, så bør værdien rettes eller fjernes. Hvis en outlier faktisk er en reel måling, så skal målingen medtages i analysen, da den repræsenterer den variation, der rent faktisk er i data. Outliers kan have en stor påvirkning på korrelationsparameteren og er derfor vigtige at behandle med omhu.

Figur 4 viser otte forskellige plots, men kun for nogle af dem giver det mening at udregne Pearson korrelationen.

Du kan nu regne øvelse 4



Figur 4. Otte forskellige xy-plots. Fire af de otte plots opfylder ikke antagelserne for bruge Pearson korrelationen.

Eksempel: Er antagelserne omkring protein- og mRNA ekspression for *EGFR* opfyldte?

I eksemplet med protein og mRNA ekspressison for *EGFR* kan vi betragte figur 2, og vi finder at:

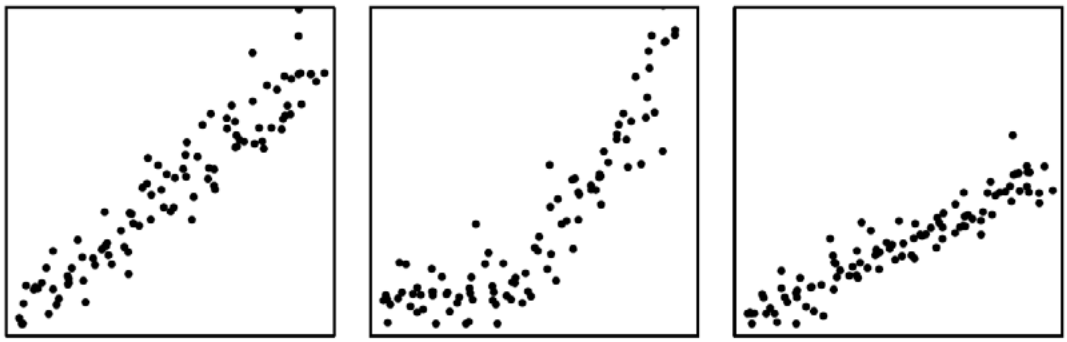
- Linearitetsantagelsen er opfyldt. Vi så en lineær sammenhæng ud fra *xy*-plottet.
- Uafhængighedsantagelsen er svær at undersøge, men vi antager, at data er samlet ind fra tilfældige personer med henblik på at være repræsentativ for den befolkning, der undersøges. Vi forventer derfor, at uafhængighedsantagelsen er opfyldt.
- Antagelsen om ingen ekstreme outliers lader også til at være opfyldt. Der er måske to observationer, som stikker lidt i øjnene nederst i højre hjørne på figur 2, men der er meget variation i data i det hele taget, og der er tilsvarende observationer i den anden side figuren, som bare indikerer, at der er en del individer, hvor de to ekspressionsniveauer afviger fra hinanden.

Da antagelserne er opfyldte, giver det mening at bruge vores udregnede Pearson korrelation på 0.80.

Øvelser

1. **Gæt korrelationen.** Brug 5-10 minutter på hjemmesiden http://guessthecorrelation.com. Hjemmesiden tester/øver din evne til at give et bud på korrelationen for forskellige sammenhænge.

2. **Korrelationsplots.** Gæt for følgende tre plots, hvilken korrelation (r_p) som hører til det specifikke plot: 0.95, 0.97 og 0.9.



3. **Udregn korrelationen**. Lad der været givet følgende 6 talpar: (1, 6), (1, 7), (3, 2), (4, -1), (6, 4) og (9, 0). Udregn korrelationen.

4. **Diskuter antagelserne**. Betragt de 8 plots i figur 4. Diskuter, hvilke plots I synes opfylder kravene til at udregne Pearson korrelationen, og hvilke, der ikke gør. Forklar, hvorfor I kommer frem til jeres resultater.

Ekstramateriale: Vurdering af korrelationskoefficienten

Indtil videre har vi ikke diskuteret usikkerheden på koefficienten for r_P, så det er svært at vurdere, om estimatet 0.8 er et stort tal eller ej. Til det formål kan vi udregne et 95% konfidensinterval for korrelationskoefficienten. Desværre følger korrelationen ikke en pæn normalfordeling, hvilket betyder, at vi skal transformere korrelationen og at udregningen bliver ret lang. Opskriften på at finde dette interval er følgende:

1. Udregn først korrelationskoefficienten, r_P.
2. Udregn

$$z = \frac{1}{2} \cdot \ln\left(\frac{1+r_P}{1-r_P}\right)$$

for at transformere korrelationen, så den er mere symmetrisk.

3. Udregn de nedre og øvre grænser:

$$\left(z - 1.96 \cdot \frac{1}{\sqrt{n-3}}; z + 1.96 \cdot \frac{1}{\sqrt{n-3}}\right),$$

hvor n er antallet af observationer.

4. Den nedre og øvre grænse skal nu tilbagetransformeres vha. formlen

$$\frac{e^{2\tilde{z}}-1}{e^{2\tilde{z}}+1},$$

hvor $\tilde{z}$ er enten den nedre eller øvre grænse udregnet i punkt 3 ovenfor. 95% konfidensintervallet er netop de tilbagetransformerede værdier.

I eksemplet med *EGFR* havde vi en korrelationskoefficient på 0.8 baseret på n=186 observationer. Et 95% konfidensinterval for korrelationskoefficienten bliver derfor udregnet på følgende måde. Først udregnes

$$z = \frac{1}{2} \cdot \ln\left(\frac{1+0.8}{1-0.8}\right) = 1.10.$$

De nedre og øvre (transformerede) grænser er derfor

$$\left(1.10 - 1.96 \cdot \frac{1}{\sqrt{186-3}}; 1.10 + 1.96 \cdot \frac{1}{\sqrt{186-3}}\right) = (0.95; 1.25).$$

Disse tilbagetransformeres til et 95% konfidensinterval for Pearsons korrelation til

$$\left(\frac{e^{2\cdot 0.95}-1}{e^{2\cdot 0.95}+1}; \frac{e^{2\cdot 1.25}-1}{e^{2\cdot 1.25}+1}\right) = (0.74; 0.85).$$

Vi er med andre ord 95% sikre på, at intervallet fra 0.74 til 0.85 indeholder den sande korrelationskoefficient i populationen. Dette kan afrapporteres som

Parameter	Estimat	95% konfidensinterval
r_P	0.80	[0.74 ; 0.85]

95% konfidensintervallet indikerer, at man er meget sikker på en positiv lineær sammenhæng, da 0 (svarende til ingen sammenhæng) er langt fra at være indeholdt i konfidensintervallet. Vi vil derfor konkludere, at der er en stærk lineær sammenhæng mellem mRNA ekspressionsniveau og proteinniveauet. Noget software kan også producere en *p* værdi for testet om, hvorvidt korrelationskoefficienten kunne være 0. Hvis *p* værdien er lille (fx. mindre end 5%), forkastes hypotesen om, at korrelationen kunne være 0.

Øvelse 5: Statistik. I et forsøg med n=19 observationer finder man en korrelation r_P= *0.45*. Udregn et 95% konfidensinterval for korrelationskoefficienten, og lav en statistisk vurdering af, om der er en lineær sammenhæng i data.

Ekstramateriale: Spearmans korrelation

Pearsons korrelation betragter *lineære* sammenhænge, men vi kan godt observere monotone sammenhænge (enten stigende eller faldende) mellem to variable, uden at sammenhængen behøver at være lineær. Spearmans korrelation bruges til at måle monotone sammenhænge.

En lineær sammenhæng er faktisk en monoton sammenhæng så Spearman korrelationen kan også benyttes til at undersøge lineære forhold. For lineære sammenhænge vil de to metoder tilnærmelsesvis give ens resultater. Dog er metoden ikke lige så præcis som Pearsons korrelation, hvis der undersøges for et lineært forhold.

Spearmans korrelationskoefficient er givet ved følgende formel:

$$r_S = \frac{\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\cdot\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\cdot\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2}}$$

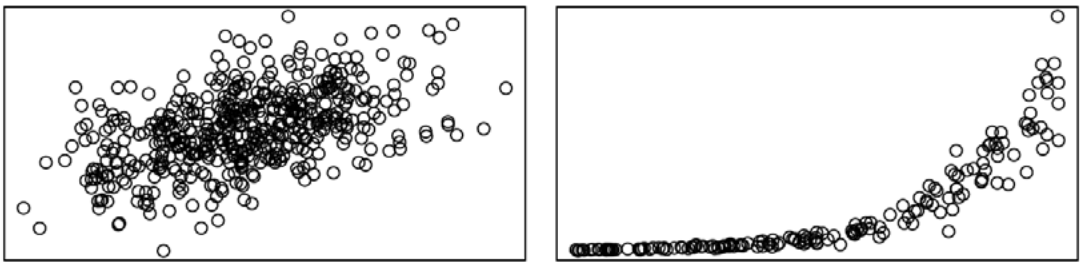
$R(x_i)$ står for rangen af x_i. Dvs. at hvert x_i får et tal afhængig af dens størrelse. Det mindste x_i får værdien 1, det næst-mindste værdien 2 osv. Tilsvarende står $R(y_i)$ for rangen af y_i. $\overline{R(x)}$ er gennemsnitsrangen af x og $\overline{R(y)}$ er gennemsnitsrangen af y. Man kan bemærke, at formlen på Pearsons korrelationkoefficient og Spearmans korrelationskoefficient grundlæggende er ens — den eneste forskel er, at man med Pearsons korrelation bruger de faktiske værdier for x og y, mens man for Spearmans korrelationskoefficient kun benytter rangen/ordningen på de tilsvarende værdier.

Et eksempel på en beregning af rangen, er vist nedenfor for udsnittet af data fra figur 1.

ID	x_i	$R(x_i)$	y_i	$R(y_i)$
TCGA-02-0003-01	-0.55	3	-1.84	3
TCGA-02-0004-01	1.65	7	-1.41	5
TCGA-02-0011-01	-1.47	1	-0.89	7
TCGA-02-0014-01	-0.75	2	-1.89	2
TCGA-02-0068-01	5.55	8	-1.92	1
TCGA-02-0069-01	-0.34	5	-0.86	8
TCGA-02-0116-01	0.29	6	-1.09	6
TCGA-02-2470-01	-0.38	4	-1.65	4

Da Spearmans korrelation kun kræver et monotont forhold kan den fange flere sammenhænge end Pearsons korrelation.

Øvelse 6: Spearman vs. Pearson. Gæt ud fra figur 5, hvilke Pearson korrelationsestimater (r_P=*0.5 og* r_P=*0.84*) og Spearman korrelationsestimater (r_s=*0.49 og* r_s=*0.99*), som hører til hvert af de to plots:

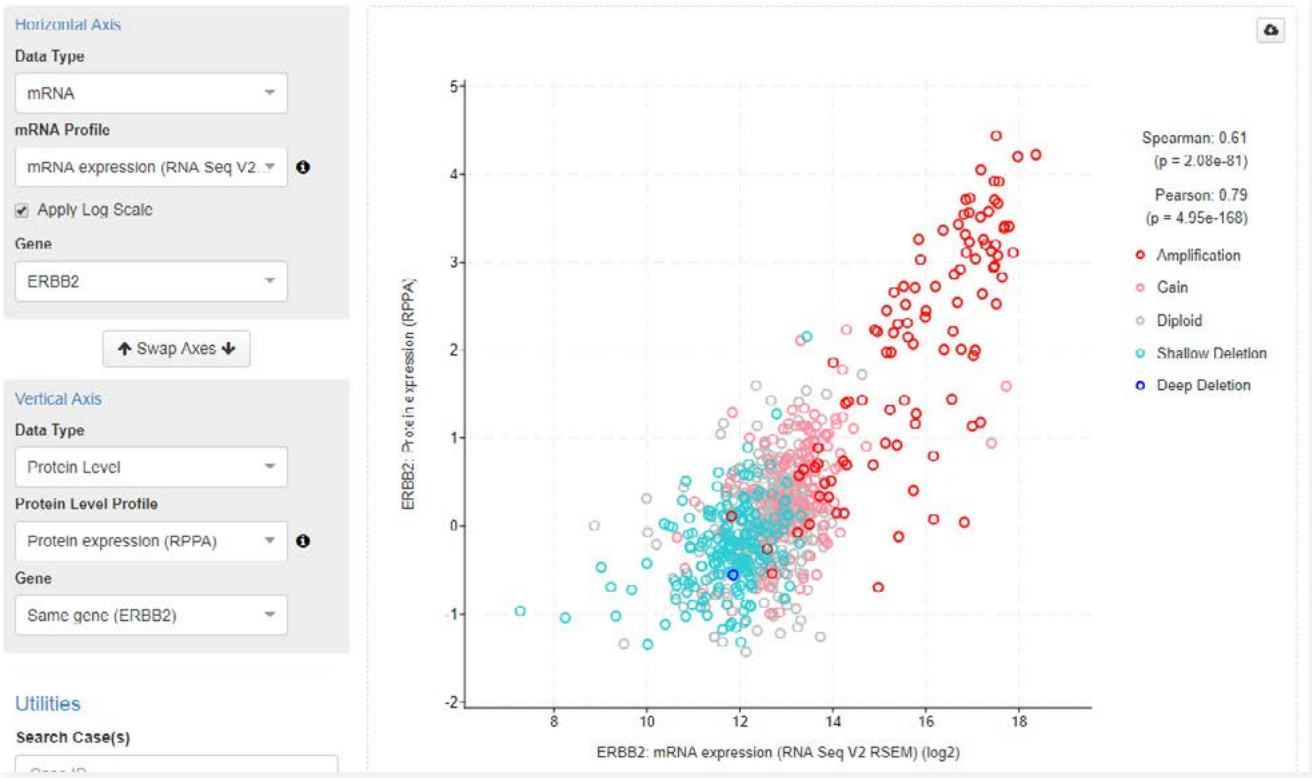


Figur 5. To *xy*-plot som viser eksempler på forhold mellem to variable med hver deres estimerede Pearson og Spearman korrelationsparameter.

Matematikøvelse 1. Brystkræft: Sammenhængen mellem mRNA ekspression og proteinniveau - en statistisk undersøgelse

Intro

Vi tager udgangspunkt i nedenstående diagram, som I allerede har set og lavet sammen med jeres Bioteknologiunderviser. Diagrammet viser som bekendt sammenhæng mellem mRNA ekspression og proteinniveau (protein level), hvor forskellige kategorier er vist med forskellig farve.



Figur 1. Sammenligning af mRNA ekspression og proteinniveau. Kategorier er vist med forskellig farve .

Der er en række statistiske muligheder i cBioPortal, men man kan også hente data ned og behandle dem et andet sted end cBioPortal.

Vi vil gøre sidstnævnte for (forhåbentlig) at gøre statistikken tydeligere og for bedre at kunne vælge, hvad vi undersøger. Resultatet af undersøgelserne skal I bringe med tilbage til jeres Biotekundersøgelse.

Den statistiske behandling - det skal I svare på

Denne del af teksten forudsætter, at I allerede har informationerne fra ovenstående graf i et Excel-regneark (Brystkræftdata_til regression.xlsx). Du kan downloade filen ved at klikke på linket. Husk at gemme filen på din egen computer. Hvis man hellere vil hente nyere data fra cBioPortal og ordne dem i et Excel regneark, kan dette gøres ved at følge bilag 1 (sidst i denne øvelse).

I regnearket er mindst tre søjler som indeholder mRNA ekspression, Proteinniveau og patient ID.

I kan nu vælge at lave den resterende del af opgaven i Excel, men I kan også vælge at overføre (udvalgte) data til jeres foretrukne dataanalyse program.

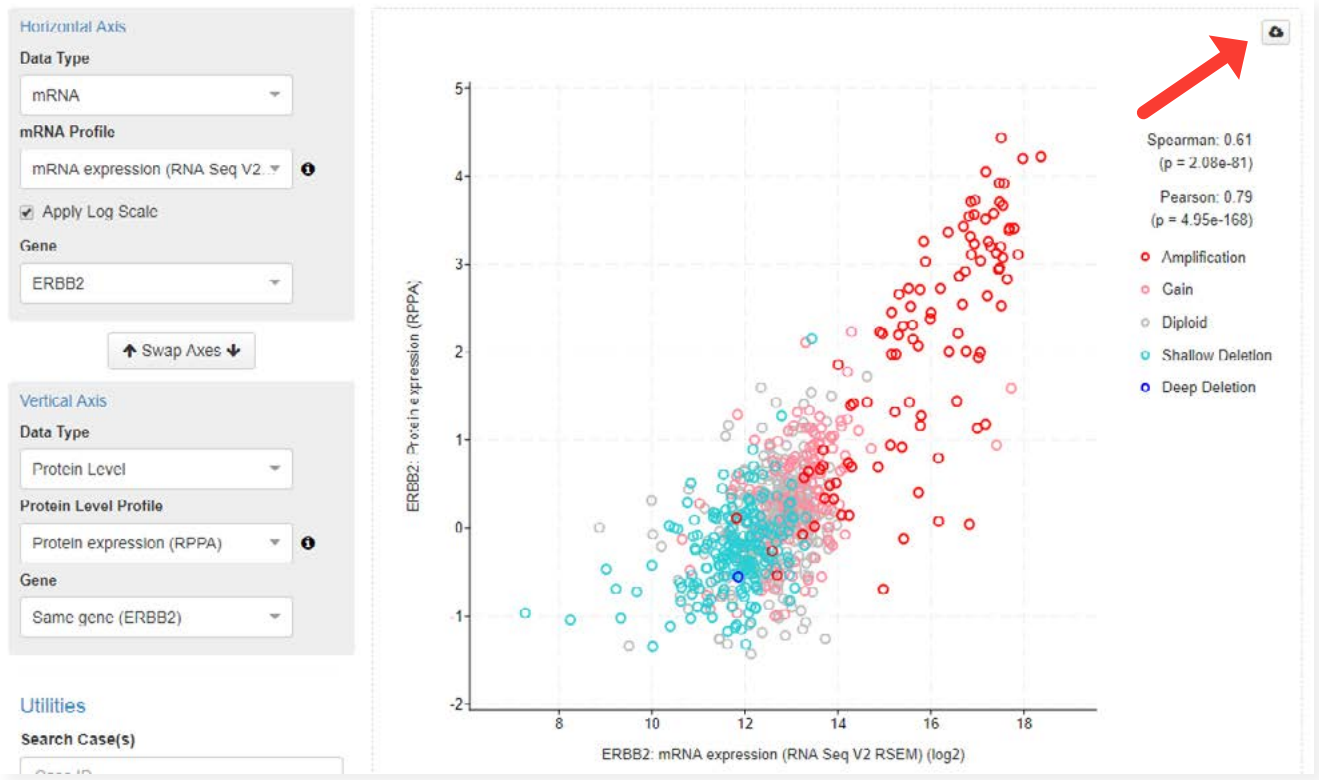
1. Lav en søjle, hvor $\log_2$ til mRNA ekspressionen beregnes.
2. Lav et diagram over alle data ($\log_2$ (mRNA) mod protein niveau) som ligner figur 1.
3. Lav lineær regression og bestem vigtige statistiske estimatorer (f.eks. korrelation). Giver dette mening hen over flere kategorier?
4. Udvælg de data fra datasættet som er i kategorien ("amplification").
5. Lav et diagram over de udvalgte data ($\log_2$ (mRNA) mod protein niveau), som ligner figur 1.
6. Lav lineær regression og bestem vigtige statistiske estimatorer (f.eks. korrelation). Giver dette mening?
7. Forklar betydningen af modellens parametre.
8. Lav et residualplot.
9. Bestem residualspredningen.
10. Kommenter modellen.
11. Gentag evt. undersøgelsen (punkt 4-10) for andre udvalgte data (der giver mening)
12. Vend nu tilbage til biotekøvelsen *"Brystkræft og genet ERBB2: ændringer i transskription og translation"* med jeres nye viden.

Bilag 1: Hente data fra cBioPortal og ordne dem i et Excel-regneark

At hente data i cBioportal

Overordnet er det ikke svært at hente data, men data kommer typisk i en tekst-fil, der er svær at læse, hvis den ikke importeres i et regneark. Det vil vi gøre i Excel (men andre programmer virker også).

Først laver vi en graf over de data, vi gerne vil hente. Vi vil gerne hente data svarende til grafen nedenfor. Kan man ikke huske, hvordan man gør, følger man vejledningen hertil i den relaterede biotekøvelse.



For at downloade data i en txt-fil (plot.txt) trykkes på skyen, som ses i øverste højre hjørne af grafen og "data" vælges.

Denne fil gemmes normalt automatisk, hvor jeres filer downloades til. Denne fil kan som tidligere fortalt hentes ind i og behandles i Excel.

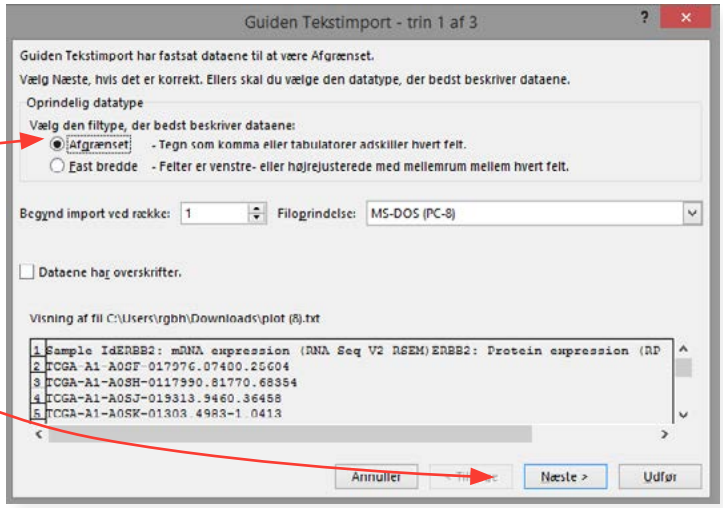
At importere en txt-fil i Excel

En tekstfil (txt-fil) importeres nemmest ved at åbne Excel og efterfølgende åbne tekstfilen fra Excel. På denne måde startes en automatisk importeringsdel i Excel. Importeringen ser nogenlunde ud som følger, men visuelt kan det se lidt anderledes ud afhængigt af Excel-version og styresystem (Win, OSX osv.).

A. Åbn tekstfilen i Excel

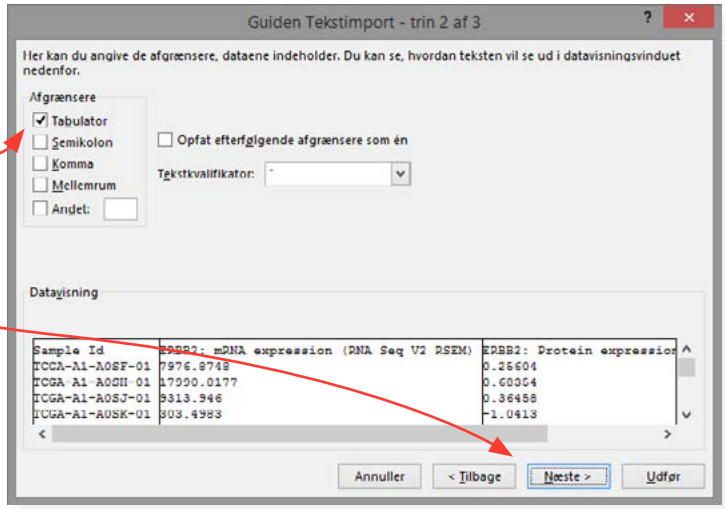
B. Sørg for at der er valgt afgrænset

C. Tryk herefter på Næste



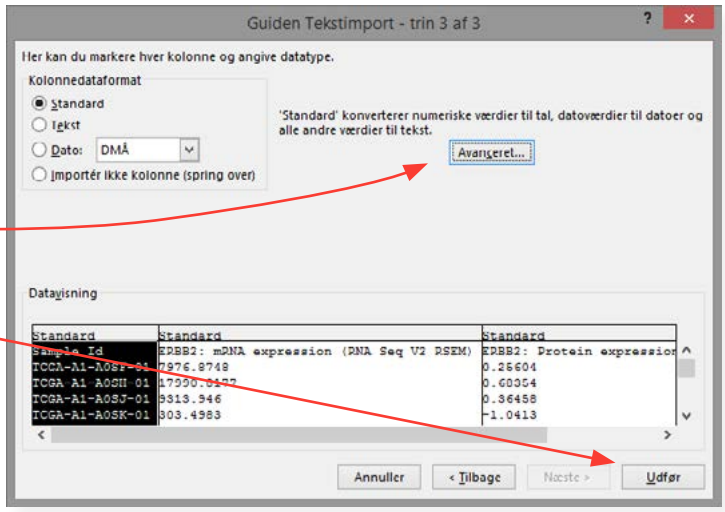
D. Sørg for at *kun* tabulator er valgt

E. Vælg næste

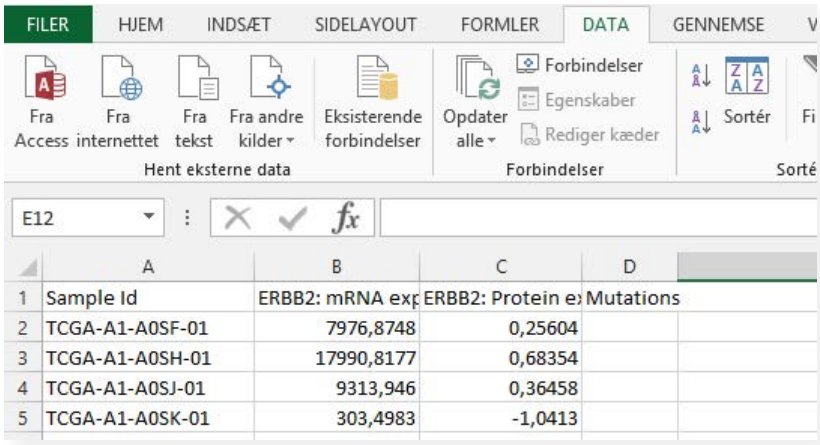


F. Sørg for at decimalseparator er . (punktum) og tusindtal-separator er , (komma) under avanceret.

G. Vælg "Udfør"



Efter vi har importeret data, vil vi nu have et regneark som ser nogenlunde ud som udsnittet til højre:

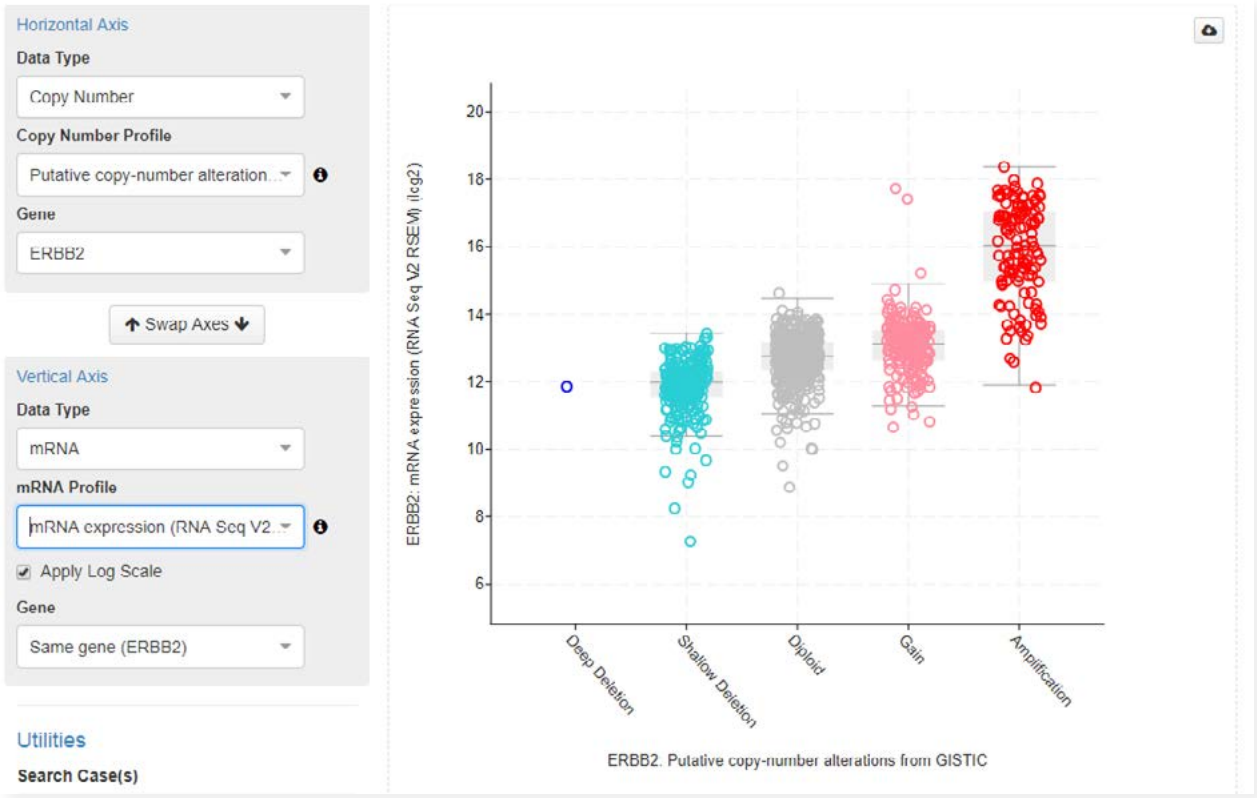


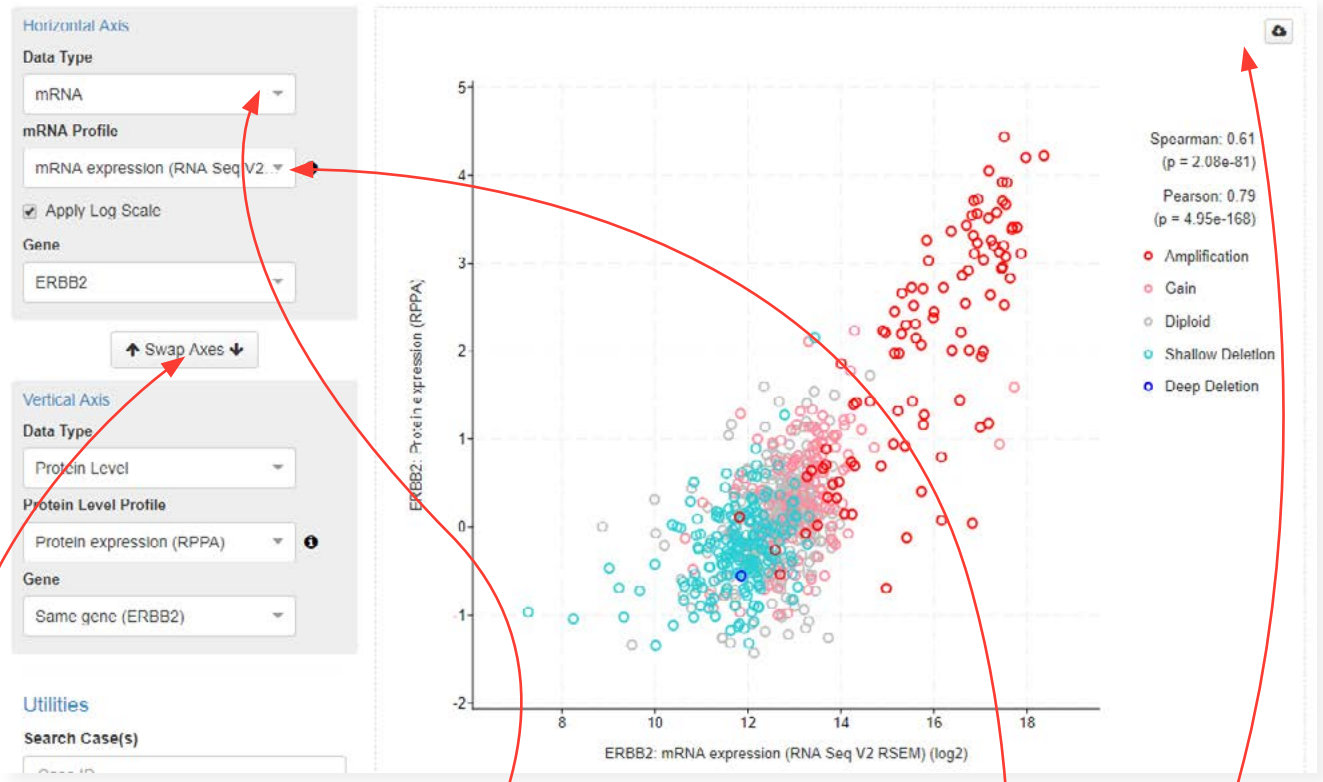
	A	B	C	D
1	Sample Id	ERBB2: mRNA exp	ERBB2: Protein e	Mutations
2	TCGA-A1-A0SF-01	7976,8748	0,25604	
3	TCGA-A1-A0SH-01	17990,8177	0,68354	
4	TCGA-A1-A0SJ-01	9313,946	0,36458	
5	TCGA-A1-A0SK-01	303,4983	-1,0413	

H. Husk at gemme filen under et passende navn, inden I går videre.

At udvælge data efter kategori

At vælge data efter kategoriseringen (her Deep Deletion, Shallow Deletion, Diploid, Gain og Amplication) kræver (desværre), at der hentes yderligere data. I denne forbindelse laver vi derfor et nyt diagram som indeholder kategoriseringen. Diagrammet, vi ønsker, ses nedenfor. Man kommer nemmest til dette diagram ved fortsat at følge vejledningen.





Diagrammet fremkommer nemmest ud fra det oprindelige diagram ved at

I. Bytte akserne

J. Vælge datatypen for den horisontale akse til "Copy Number" og profilen til "Putative copy-number..."

K. Hente data ned som en tekstfil (txt-fil)

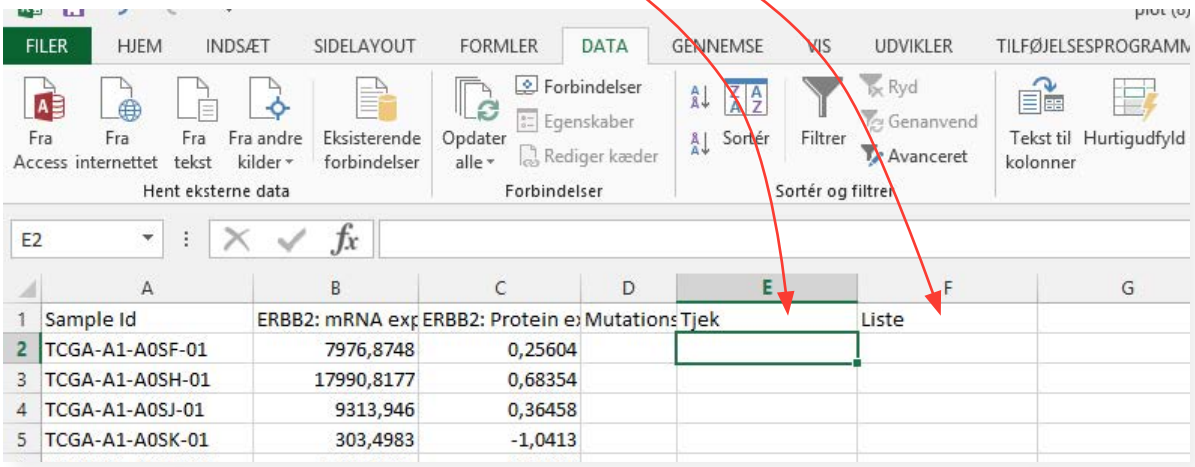
L. Importere tekstfilen i Excel (på samme måde som tidligere) og huske at gemme.

Den ene Excel-fil indeholder nu data om mRNA og Proteinniveau, mens den anden indeholder information om mRNA og kategorisering. Vi løber dog ind i et problem med, at de to dataserier ikke nødvendigvis indeholder de samme patienter, idet cBioPortal hele tiden justerer for at få flest personer med. Vi er derfor nødt til at sortere os frem, hvilket er lidt indviklet.

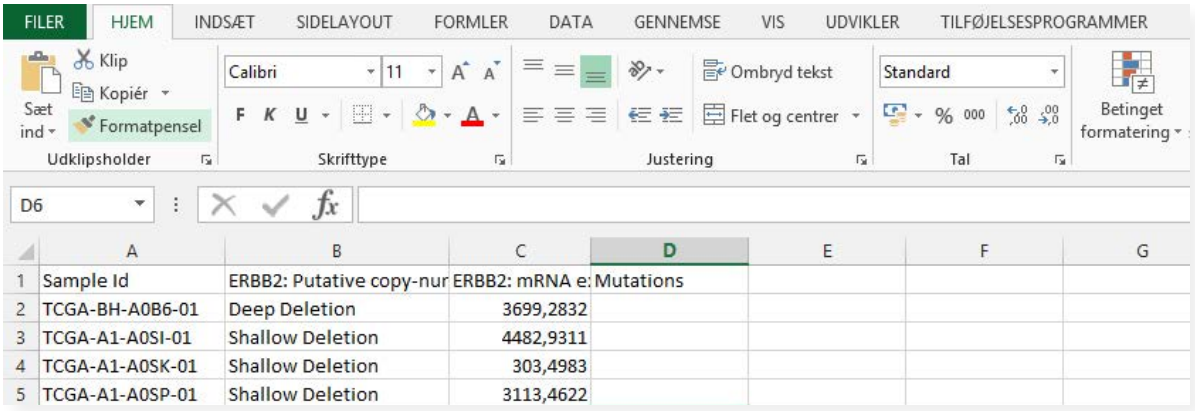
Sorteringen kan udføres ved at følge denne lidt indviklede procedure

A. Excelfilen med data om mRNA og Protein niveau åbnes (hvilket ligner nedenstående figur).

B. Der tilføjes en overskrift i celle E1 og F1. Titlerne er ikke vigtige, men de kunne være Tjek og Liste

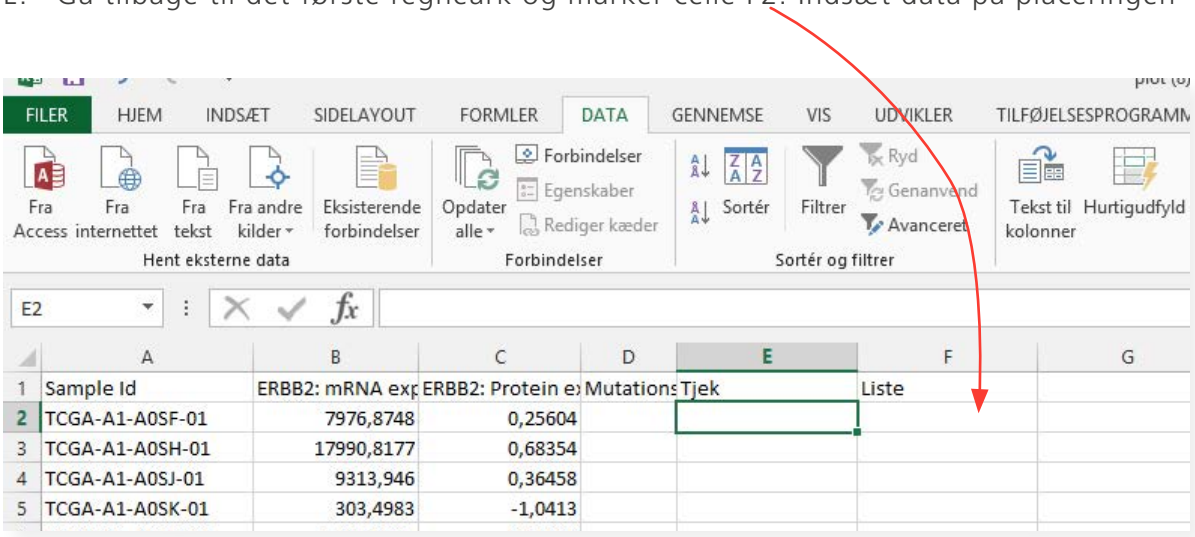


C. Den anden Excel-fil åbnes (den med kategorierne)



D. "Sample Id" for alle patienter, vi vil vælge, markeres. Vi vælger sample Id for alle patienter, som i kolonne B (ERBB2: Putative...) har beskrivelsen "amplification". Kopier data.

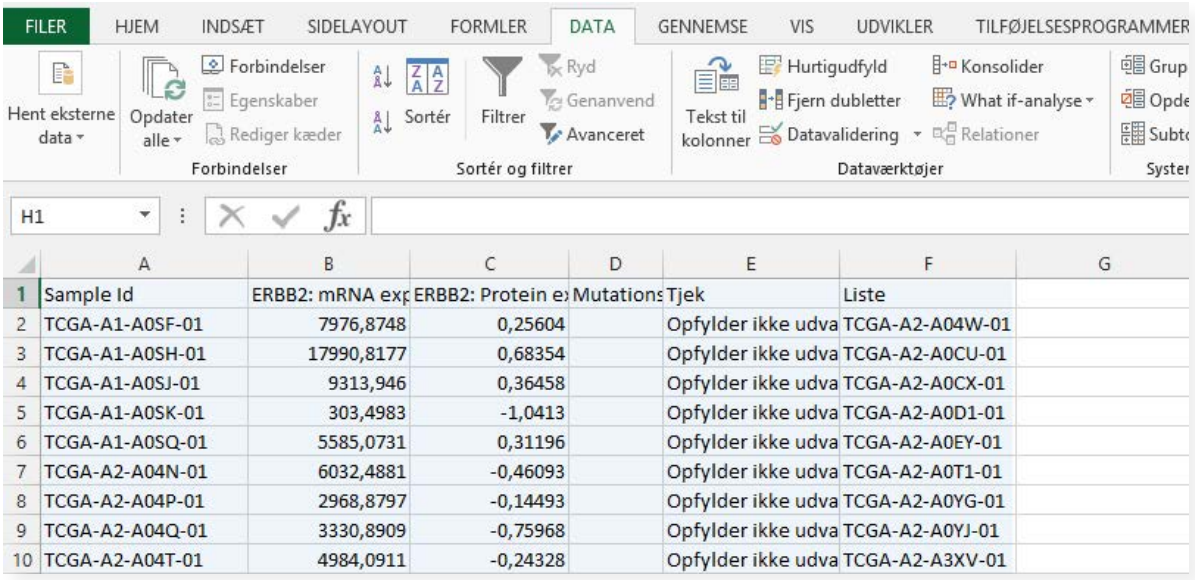
E. Gå tilbage til det første regneark og marker celle F2. Indsæt data på placeringen



Nu mangler vi bare at udvælge (også selvom man har valgt alle kategorierne)

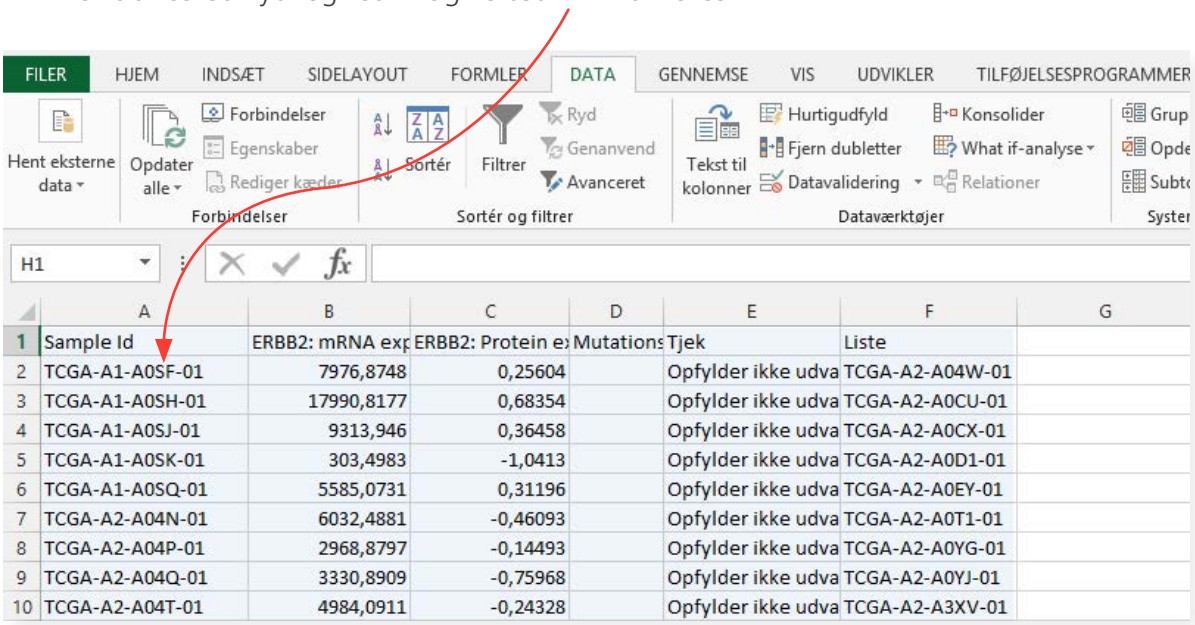
F. Hvis man har en dansk udgave af Excel, skives der i celle E2 helt nøjagtigt
=HVIS(ER.TAL(SAMMENLIGN(A2;$F$2:$F$1000;0));"OK";"Opfylder ikke udvalget")
men hvis man har en engelsk udgave af Excel, skives der i celle E2 helt nøjagtigt
=IF(ISNUMBER(MATCH(A2;$F$2:$F$1000;0));"OK";"Opfylder ikke udvalget")

Resultatet bliver følgende:



G. Alle data markeres og kopieres

H. Der åbnes et nyt regneark og feltet A1 markeres.



	A	B	C	D	E	F	G
1	Sample Id	ERBB2: mRNA exp	ERBB2: Protein e	Mutations	Tjek	Liste	
2	TCGA-A1-A0SF-01	7976,8748	0,25604		Opfylder ikke udva	TCGA-A2-A04W-01	
3	TCGA-A1-A0SH-01	17990,8177	0,68354		Opfylder ikke udva	TCGA-A2-A0CU-01	
4	TCGA-A1-A0SJ-01	9313,946	0,36458		Opfylder ikke udva	TCGA-A2-A0CX-01	
5	TCGA-A1-A0SK-01	303,4983	-1,0413		Opfylder ikke udva	TCGA-A2-A0D1-01	
6	TCGA-A1-A0SQ-01	5585,0731	0,31196		Opfylder ikke udva	TCGA-A2-A0EY-01	
7	TCGA-A2-A04N-01	6032,4881	-0,46093		Opfylder ikke udva	TCGA-A2-A0T1-01	
8	TCGA-A2-A04P-01	2968,8797	-0,14493		Opfylder ikke udva	TCGA-A2-A0YG-01	
9	TCGA-A2-A04Q-01	3330,8909	-0,75968		Opfylder ikke udva	TCGA-A2-A0YJ-01	
10	TCGA-A2-A04T-01	4984,0911	-0,24328		Opfylder ikke udva	TCGA-A2-A3XV-01	

I. Data indsættes ved at højreklikke (to-fingerklikke) på cellen A1, vælge indsæt special og vælge værdier

J. Alle data markeres, og man vælger sortér under data. Data sorteres efter kolonne E (Tjek).

K. De relevante data står nu øverst, og de resterende data kan let slettes. Regnearket er nu klar til at lave den egentlige øvelse.

Chi-i-anden og sammenligning af sandsynligheder

Anne Lyngholm, Bjarke Hansen og Claus Ekstrøm | November 2019

Det skal vi lære

Når I har været gennem materialet, skal I kunne svare på følgende:

- Hvad kan vi bruge et chi-i-anden test til?
- Hvad kan man konkludere ud fra et chi-i-anden test?
- Hvad fortæller fortegnet på odds-ratio (OR) os?
- Hvordan anvendes ensidede tests til at teste for mutual exclusivity/co-occurrence?

Introduktion

Vi ved fra matematikundervisningen, at binomialfordelingen kan bruges til at modellere og teste hypoteser omkring en sandsynlighed, når man har *n* observationer, og man for hver observerer netop et af to mulige udfald. Her skal vi se på, hvad vi gør, hvis vi vil sammenligne sandsynlighederne fra to grupper, eller hvis vi vil undersøge, om der er sammenhæng mellem to variable, hvor der er to mulige udfald for hver variabel. I begge tilfælde kan man benytte et chi-i-anden test til at vurdere, om de to sandsynligheder er ens eller om der er sammenhæng mellem de to variable.

Eksempel: Onkogenerne *KRAS* og *EGFR*

For nogle kræfttyper er det interessant at undersøge, hvorvidt visse kræftgener sjældent er aktiverede samtidig, som det er tilfældet med *KRAS* eller *EGFR*. Der findes studier, som indikerer, at disse to onkogener er syntetisk letale, hvilket betyder, at tumoren ikke kan tåle at begge gener er muteret i den samme kræftcelle. Vi vil starte med at undersøge, om sandsynligheden for at *EGFR* er muteret er uafhængig af, om *KRAS* er muteret - det vil sige, at de arbejder uafhængigt af hinanden. Et udsnit af data, som kan benyttes til at undersøge hypotesen, er vist i figur 1. For at få et større overblik opsummeres det fulde datasæt som vist i tabellen nedenfor, så bl.a. hyppigheden af muteret/ikke-muteret for de to onkogener er tydelig.

	A	B	C
1	**Patient _ID**	**KRAS**	**EGFR**
2	1	Ja	Ja
3	2	Ja	Nej
4	3	Nej	Nej
5	4	Nej	Nej
6	5	Ja	Nej
7	6	Nej	Ja
8	7	Nej	Ja
9	8	Ja	Nej
10	9	Nej	Ja
11	10	Ja	Ja

Figur 1. Udsnit af data, der viser om de enkelte onkogener er aktive, det vil sige enten muteret eller amplificeret.

	EGFR: Muteret	*EGFR:* Ikke-muteret	Total
KRAS: Muteret	5 (6.1%)	77 (93.9%)	82 = 5 + 77
KRAS: Ikke-muteret	35 (23.6%)	113 (76.4%)	148 = 35 + 113
Total	40 = 5 + 35	190 = 77 + 113	230 = 40 + 190 = 82 + 148

I tabellen tilhører hver person netop én kombination af *KRAS* status (muteret/ikke-muteret) og *EGFR* status (muteret/ikke-muteret). Af de i alt 148 personer har 5 personer begge gener muterede, og blandt de 82 personer, hvor *KRAS* er muteret, er der 5 personer (6.1%), der også er muterede for *EGFR*. I dette eksempel er der to variable af interesse (de to onkogener), som hver er inddelt i to kategorier (muteret/ikke-muteret). Generelt må antallet af kategorier for den enkelte variabel godt være større end to og der behøver ikke at være det samme antal kategorier for de to variable af interesse for at undersøge, om der er en sammenhæng mellem de to variable.

Hvis de to inddelinger i tabellen (*KRAS* og *EGFR* ovenfor) var uafhængige, så skulle det være sådan, at hvis man vidste, om *KRAS* var muteret, så ville vi *ikke* være bedre i stand til at gætte, om *EGFR* var muteret eller ej. Hvis θ_A er sandsynligheden for at *EGFR* er muteret blandt de personer, der har *KRAS* muteret, og θ_{IA} er sandsynligheden for at *EGFR* er muteret blandt de personer, hvor *KRAS ikke* er muteret, så vil uafhængigheden mellem de to gener svare til hypotesen om at

$$H_0 : \theta_A = \theta_{IA}.$$

Hypotesen betyder, at uanset om *KRAS* er muteret eller ej, så er der samme sandsynlighed for, at *EGFR* er muteret. De har altså ikke indflydelse på hinanden.

I tabellen ovenfor spørger vi altså, om de 6.1% og 23.5% sandsynlighed for muteret *EGFR* i de to grupper af *KRAS* i bund og grund kunne stamme fra samme (underliggende) risiko i de to grupper.

Hvis hypotesen er sand, er der samme sandsynlighed for mutation af *EGFR* i begge grupper af *KRAS*, og det bedste bud på denne fælles sandsynlighed må være 40/230 = 17.39%, idet der i alt er netop 40 patienter med muterede *EGFR* gener ud 230 patienter. Det betyder også, at vi ville have forventet, at

$$\text{Forventet antal} = \begin{cases} \frac{40}{230} \cdot 82 = 14.26 \text{ i gruppen, hvor } \textit{KRAS} \text{ er muteret } (82 \text{ patienter}) \\ \frac{40}{230} \cdot 148 = 25.74 \text{ i gruppen, hvor } \textit{KRAS} \text{ ikke er muteret } (148 \text{ patienter}) \end{cases}$$

antallet af observationer med *EGFR* muteret havde været

Vi kan sammenfatte de observerede og de forventede data i følgende to tabeller.

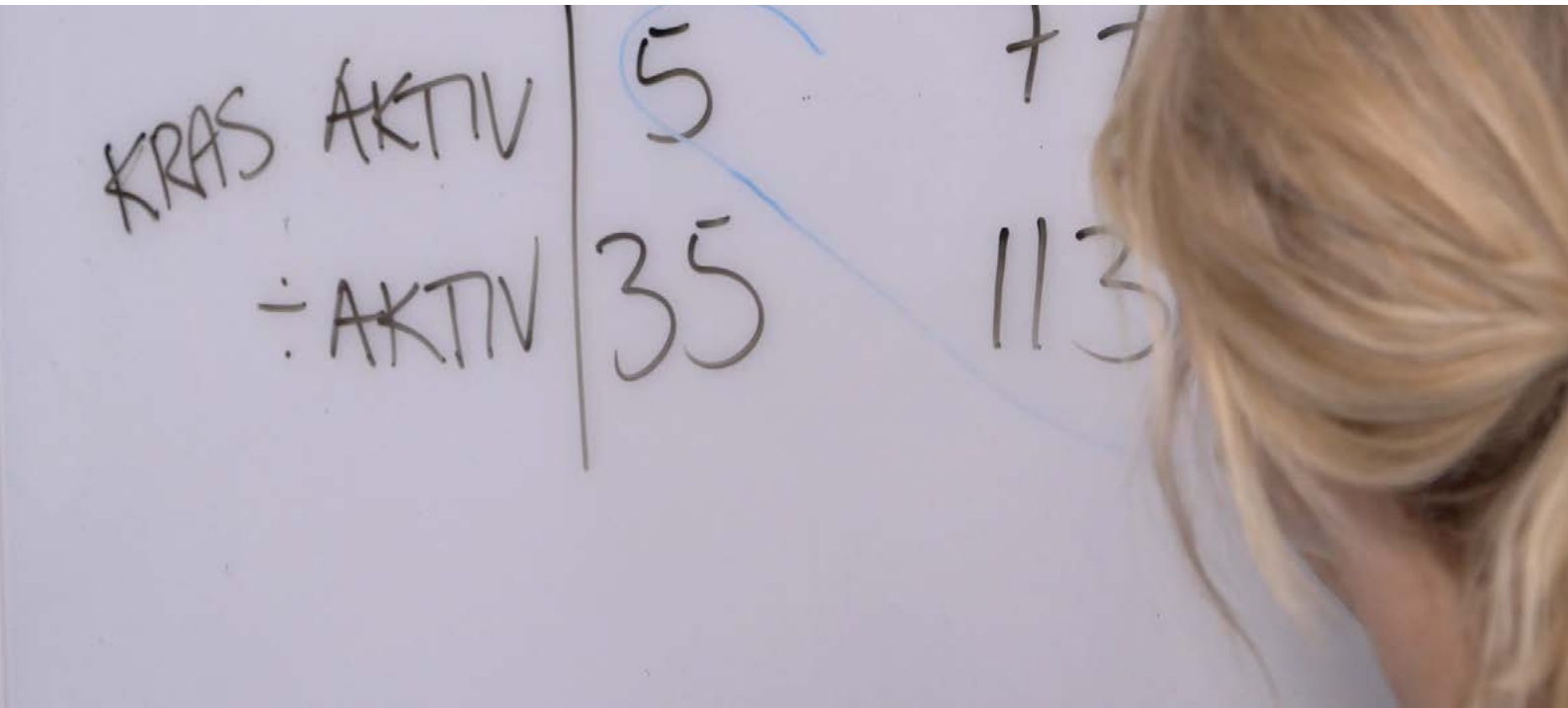
Observerede antal:

	EGFR: Muteret	*EGFR:* Ikke-muteret	Total
KRAS: Muteret	**5**	**77**	82
KRAS: Ikke-muteret	**35**	**113**	148
Total	40	190	230

Forventede antal:

	EGFR: Muteret	*EGFR:* Ikke-muteret	Total
KRAS: Muteret	40/230 · 82 = **14.26**	190/230 · 82 = **67.74**	82
KRAS: Ikke-muteret	40/230 · 148 = **25.74**	190/230 · 148 = **122.26**	148
Total	40	190	230

Umiddelbart ligner de forventede og observerede observationer ikke hinanden helt ud fra tabellerne ovenfor: 5 er eksempelvis noget mindre end 14.26, og 77 er noget større end 67.74, men vi har brug for en objektiv måde at vurdere, om de to talsæt ligger tæt på hinanden eller ej, og her kommer chi-i-anden-testet ind.



Chi-i-anden-testet for sammenhæng i antalstabeller

Chi-i-anden teststørrelsen måler, om værdierne (observerede og forventede) er tæt på hinanden. Hvis de er det, kan vi ikke afvise, at de kunne komme fra samme fordeling. Hvis de er langt fra hinanden, afviser vi hypotesen om, at observerede værdier stammer fra samme fordeling. Chi-i-anden teststørrelsen er defineret som:

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - F_i)^2}{F_i} = \frac{(O_1 - F_1)^2}{F_1} + \frac{(O_2 - F_2)^2}{F_2} + \cdots + \frac{(O_K - F_K)^2}{F_K},$$

hvor *O* står for Observeret, *F* for Forventet og *K* for antallet af Kategorier (hvilket svarer til antallet af forskellige kombinationer af inddelinger). Hvis de observerede data stemmer overens med de forventede data, vil tællerne være tæt på 0, og teststørrelsen bliver tæt på nul. Hvis observerede og forventede værdier er forskellige bliver teststørrelsen stor. Jo større teststørrelse jo dårligere stemmer de forventede og observerede overens.

Eksempel: Udregning af teststørrelsen for sammenhæng mellem *KRAS* og *EGFR*

	O	*F*	$(O - F)$	$(O - F)^2$	$(O - F)^2 / F$
Muteret *KRAS*, muteret *EGFR*	5	14.26	-9.26	85.75	6.01
Muteret *KRAS*, ikke-muteret *EGFR*	77	67.74	9.26	85.75	1.27
Ikke-muteret *KRAS*, muteret *EGFR*	35	25.74	9.26	85.75	3.33
Ikke-muteret *KRAS*, ikke-muteret *EGFR*	113	122.26	-9.26	85.75	0.70

Lad os udregne chi-i-anden teststørrelsen i eksemplet med fordelingen af *KRAS* og *EGFR*. Mellemregningerne er vist i tabellen ovenfor.

Som vi kan se i tabellen, betyder størrelsen af den forventede værdi meget. Det straffes hårdt, at der er en afstand på cirka 9 personer i forskel på den observerede og forventede værdi for både at have muteret *KRAS* og *EGFR*. Her er den forventede værdi cirka 14 personer. Det straffes mindre hårdt, at der er cirka 9 personer i forskel på det forventede og observerede antal personer med ikke-muteret *KRAS* og *EGFR*. Her er den forventede værdi cirka 122 personer. Selve teststørrelsen bliver

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - F_i)^2}{F_i} = 6.01 + 1.27 + 3.33 + 0.70 = 11.31$$

Talværdien 11.31 er svær at forholde sig til og vi har brug for en "målestok" for at sige om tallet er stort eller ej. Med én frihedsgrad (se næste afsnit) er sandsynligheden for at observere 11.31 eller mere 0.15%, hvis hypotesen om samme sandsynlighed for mutation af *EGFR* er sand. Med andre ord vil der kun være 0.15% sandsynlighed for at observere en tabel så langt fra den forventede, hvis hypotesen var korrekt. *p*-værdien er derfor 0.15%, og vi forkaster derfor hypotesen om, at sandsynligheden for muteret *EGFR* er ens for de to grupper af *KRAS*.

Antallet af frihedsgrader

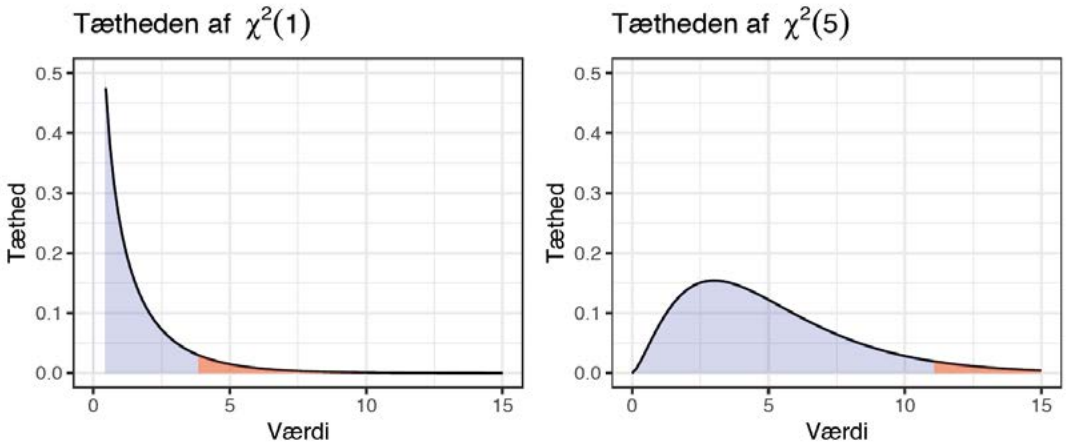
Chi-i-anden teststørrelsen følger approksimativt en fordeling kaldet chi-i-anden fordelingen (også skrevet som $\chi^2(df)$-fordelingen, hvor χ er det græske bogstav chi). df står for ændringen af frihedsgrader, der angiver, hvor stor en forskel i fleksibilitet, der er mellem modellen under hypotesen og den alternative hypotese.

Formlen for udregning af frihedsgrader (degrees of freedom) i en tabel med *r* rækker og *k* søjler er

$$\text{df} = (r-1)\cdot(k-1),$$

hvor *r* svarer til antallet af kategorier for den ene variabel (antallet af rækker i tabellen), mens *k* svarer til antallet af kategorier i den anden variabel (antallet af kolonner/søjler i tabellen).

Det er vigtigt at vide, at chi-i-anden testet er afhængig af antallet af frihedsgrader, da $\chi^2(df)$ fordelingen ændrer form afhængig af df. Når formen ændres, ændres også grænseværdien (også kaldet den kritiske værdi) for X^2, hvor vi ikke kan afvise hypotesen. I figur 2 er denne grænseværdi afspejlet ved overgangen fra de blå til de røde områder. Får man en teststørrelse, X^2, der er *større* end den kritiske grænseværdi, så vil man have en *p*-værdi mindre end 5%, og man vil forkaste hypotesen. Bemærk, at den kritiske værdi af X^2 bliver større (for et bestemt signifikansniveau), når antallet af frihedsgrader øges.



Figur 2. Tætheder for χ^2-fordelingen med 1 og 5 frihedsgrader. I begge figurer svarer det røde areal til netop 5% af hele fordelingen. Den kritiske værdi for en χ^2-fordeling bliver derfor den *x*-værdi, hvor det røde område starter.

Du kan nu regne øvelse 2

Eksempel: *KRAS* og *EGFR*

I eksemplet med effekten af muteret/ikke-muteret *KRAS* på sandsynligheden for muteret *EGFR* er antallet af frihedsgrader lig $\mathrm{df} = (2-1)\cdot(2-1) = 1$, da antallet af rækker er lig to og antallet af søjler er to. Med en frihedsgrad er grænsen for et chi-i-anden test 3.84, hvis man tester med et signifikansniveau på 5%. Denne grænse kan også aflæses på figur 2. Da vi har $X^2 = 11.31$ med én frihedsgrad forkaster vi hypotesen om, at sandsynligheden af muteret *EGFR* er uafhængig af muteret/ikke-muteret *KRAS*, da teststørrelsen er meget større end grænsen på 3.84. Der lader altså til at være en sammenhæng mellem, hvorvidt *KRAS* er muteret og *EGFR* er muteret.

Mutually exclusive og co-occurrence

Vi har netop set, at der er forskel på sandsynligheden for muteret *EGFR* afhængig af om *KRAS* er muteret eller ikke-muteret. Når de to gener ikke er uafhængige af hinanden ønsker vi at se på, om de kunne være enten *mutually exclusive* (lav sandsynlighed for at de begge er muterede eller ikke-muterede på samme tid) eller *co-occuring* (stor sandsynlighed for at begge er muterede eller ikke-muterede på samme tid).

Man kan ud fra data vurdere, om generne kunne være mutually exclusive eller co-occuring ved at se på forholdet mellem antallet af observationer blandt de fire tal i tabellen. For nemheds skyld kalder vi de fire værdier i tabellen (se tabellen under) for *A*, *B*, *C* og *D*. *A* svarer til antallet af personer, hvor begge gener er muterede, *B* svarer til antallet af tilfælde hvor *KRAS* er muteret men *EGFR* ikke er osv.

	EGFR: Muteret	*EGFR*: Ikke-muteret
KRAS: Muteret	*A*	*B*
KRAS: Ikke-muteret	*C*	*D*

Vi vil bruge følgende formel til at beskrive forholdet mellem de fire kombinationsmuligheder:

$$OR = \frac{A \cdot D}{C \cdot B},$$

hvor vi ser på (produktet) af antallene i tabellens ene diagonal divideret med (produktet) af antallet i den anden diagonal. Husk, at antallene *A* og *D* netop svarer til de observationer, hvor man har den samme aktivitet af *KRAS* og *EGFR* - enten begge muterede eller begge ikke-muterede. *OR* kaldes også odds-ratio, og den kan have værdier fra 0 til uendelig. Vi vil bruge *OR* til at undersøge, om *KRAS* og *EGFR* er mutually exclusive eller co-occuring.

Mutually exclusive. Mutually exclusive hentyder til, at vi primært forventer, at netop et af de to gener er muterede. I forhold til tabellen vil det betyde, at vi forventer, at antallet af tilfælde hvor *KRAS* og *EGFR* begge er muterede eller ikke-muterede er relativt mindre end de andre kombinationsmuligheder. Det svarer til, at *A* og *D* i tabellen er meget mindre end de øvrige to værdier. For odds-ratio svarer mutually exclusive til at tælleren bliver mindre end nævneren. En værdi af $OR<1$ er derfor et tegn på mulig mutual exclusivity. Vi kan teste hypotesen om mutual exclusivity formelt ved brug af et ensidet chi i anden test. Det vises i eksemplet nedenfor.

Co-occurence. Hvis man derimod forventer, at de to gener er co-occuring, så skal sandsynligheden for, at begge gener enten er muterede eller ikke-muterede være stor. Det svarer til, at *B* og *C* må være relativt små sammenlignet med *A* og *D*, og tilsvarende at $OR>1$. En værdi over 1 af *OR* er derfor tegn på co-occurence af generne, og vi viser nedenfor, hvordan vi tester dette ved hjælp af et ensidet chi-i-anden test.

Eksempel: test for mutal exclusivity eller co-occurrence mellem *KRAS* og *EGFR*

KRAS og *EGFR* lader til at være mutually exclusive, da tabellen viser meget få tilfælde, hvor begge gener er muterede på samme tid. Vi beregner *OR* for *KRAS* og *EGFR*, hvilket giver

$$OR = \frac{A \cdot D}{C \cdot B} = \frac{5 \cdot 113}{35 \cdot 77} = 0.21$$

Da $OR<1$ indikerer det, at de to gener er mutually exclusive, men vi er nødt til lave et formelt test, for at vurdere, om de 0.21 er langt nok væk fra 1 til, at det ikke blot skyldes tilfældigheder.

For at teste hypotesen om mutual exclusivity benytter vi os af chi-i-anden testet, som vi allerede har udregnet. Her fik vi tidligere, at det tosidede test havde en p-værdi på 0.0015, hvilket svarer til, at vi forkastede hypotesen $OR=1$, når vi sammenligende med, at $OR\neq1$, da både $OR<1$ og $OR>1$ testes.

Et ensidet test er stærkere end et tosidet, da man kun er interesseret i at afvise en af de to betingelser (enten $OR<1$ eller $OR>1$). Når man ikke har brug for at medtage den ene halvdel af testen (for mutual exclusivity er vi for eksempel ikke interesseret i påstanden om, at $OR>1$), så udføres det ensidede test ved at halvere p-værdien fra det tosidede test, når *OR* stemmer overens med den retning, som hypotesen påstår. For mutual exclusivity svarer det til at teste nulhypotesen

$$H_0 : OR = 1 \text{ mod alternativet at } OR < 1$$

I eksemplet med *KRAS* og *EFGR* finder vi, at *OR*=0.21, hvilket stemmer overens med retningen for mutual exclusivity. Hvis vi skal teste den ensidede hypotese om mutual exclusivity bliver *p*-værdien derfor 0.0015/2=0.00075. Vi forkaster altså hypotesen om ingen sammenhæng, og ser, at data lader til at stemme overens med mutual exclusivity.

Havde vi observeret en *OR*, som *ikke* var i overensstemmelse med vores hypotese (fx. hvis *OR* havde været større end 1), så er data tydeligvis ikke i modstrid med nulhypotesen om uafhængighed. I det tilfælde konkluderer man bare, at data ikke er i modstrid med nulhypotesen, og at *p*-værdien er langt større end 0.05. Hvis vi i stedet havde været interesseret i hypotesen om co-occurrence, *OR*>1, så ville vi forkaste denne hypotese, da vi i vores eksempel har beregnet *OR*=0.21<1.

Inden for BioTek er der tradition for, at *OR* ofte afrapporteres logaritme-transformeret (med base 2):

$$\log_2(OR) = \log_2\left(\frac{A \cdot D}{C \cdot B}\right) = \log_2\left(\frac{5 \cdot 113}{35 \cdot 77}\right) = \log_2(0.21) = -2.254$$

Transformationen gør, at værdier under 0 er et tegn på mutual exclusivity, mens værdier over 0 er et tegn på co-occurence. Vi får samme konklusion, hvad enten vi transformerer eller ej. Det eneste som ændrer sig er, om man sammenligner med 1 (utransformeret) eller 0 (logaritme-transformeret). Grunden til at totalslogaritmen bliver benyttet er, at hver hele værdi af ændringen af OR svarer til en fordobling af odds. Når vi får -2.25 betyder det, at der er mere end 4 gange (2 gange en fordobling) flere observationer i diagonalen af tabellen (værdierne *A* og *D*), end der er udenfor.

Øvelser

1. **Mendels ærteforsøg**. I denne øvelse skal I forholde jer til en undersøgelse som inkluderer to variable (form og farve) med to grupper hver (runde og rynkede samt grønne og gule). Eksemplet kommer fra en af genetikkens grundlæggere Gregor Mendel som undersøgte forskellige hypoteser om ærteplanter. Han opsamlede 556 planter til sin undersøgelse. De 556 planter fordelte sig på følgende kategorier:

Phenotype	Runde og gule	Runde og grønne	Rynkede og gule	Rynkede og grønne
Antal	315	108	101	32

   En specifik hypotese lød på, om farven af ærteplanten var uafhængig af formen, og den vil vi nu teste.

   a. Opstil en 2x2 tabel af data fra ærteplanterne ud fra overstående undersøgelse, hvor man ønsker at sammenligne ærteplantens farve med formen.

   b. Beskriv først hypotesen som Gregor Mendel ville undersøge i ord. Beskriv den derefter ved brug af sandsynlighedsparametre.

   c. Beregn de forventede værdier under hypotesen og opskriv dem i en 2x2 tabel. Overvej ud fra tabellen (uden at beregne X^2), om du forventer at hypotesen kan afvises.

   d. Beregn nu chi-i-anden teststørrelsen (X^2). Hvis værdien er over 3.84 afvises hypotesen. Kan du afvise hypotesen? Beskriv din konklusion i ord.

2. ***P*-værdien i en χ^2fordeling.** Et chi-i-anden test med 1 frihedsgrad har en kritisk værdi på 3.84. Hvis teststørrelsen bliver større end 3.84 vil *p*-værdien tilsvarende blive mindre end 5%. Vi kan selv overbevise os om, at 3.84 er den kritiske værdi ved at finde den værdi i en chi-i-anden fordeling, hvor 95% af sandsynlighedsmassen er til venstre (og dermed tilsvarende, at 5% er til højre).

   Dette kan f.eks. gøres via Excel, hvor funktionen CHISQ.INV(0,95; 1) oplyser grænsen (bemærk at i de danske versioner af Excel hedder funktionen CHI2.INV(0,95; 1)).

   a. Brug funktionen til at bekræfte, at grænsen for en χ^2fordeling med en frihedsgrad er 3.84.

   b. Brug funktionen i Excel til at finde de kritiske grænser for en fordeling med henholdsvis 2, 4 og 11 frihedsgrader.

3. **Mutual Exclusive og Co-occurence**. I skal i denne opgave bruge følgende tre tabeller:

Tabel 1 (Mutation af gen X):

	Somatisk mutation	Ingen somatisk mutation
Germline mutation	18	73
Ingen germline mutation	10	408

Tabel 2 (Vaccination og Influenza):

	Influenza	Ikke Influenza
Vaccine	20	100
Placebo	80	140

Tabel 3 (Farve og form på ærteplanter):

	Runde	Rynkede
Gule	315	101
Grønne	108	32

Til hver tabel er der lavet et chi-i-anden test som gav henholdsvis følgende *p*-værdier: <0.0001, 0.0002, 0.8208.

   a. Beregn odds ratio (OR) for alle tre tabeller.

   b. Baseret på om OR var over/under 1 vurder da, om der lader til at være mutual exclusivity eller co-occurence.

   c. Beregn det ensidede test for alle tre hypoteser.

4. **Tynde tabeller.** Når antallet af observationer i en krydsstabel er lille, kan man ikke være sikker på, at værdien af teststørrelsen er præcis nok til, at man kan sammenligne værdien med en χ^2fordeling. En grov tommelfingerregel plejer at være, at det **forventede** antal i hver celle skal være større end 5 for, at der er nok data til, at teststørrelsen har stabiliseret sig.

   Betragt nedenstående tabel mellem to gener (*CDKN2A og RB1*).

   a. Udregn den forventede tabel. Kan man have tillid til *p*-værdierne udregnet på baggrund af tabellen?

   b. Udregn odds ratio (*OR*)? Hvad er problemet med at lave denne udregning?

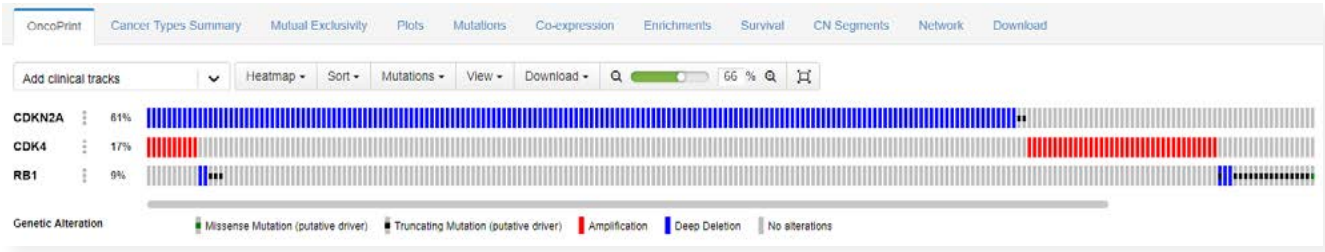
   c. Når nogle udfald er meget sjældne, bliver odds ratio ustabil, fordi man umiddelbart vil estimere, at en hændelse aldrig vil forekomme. Hvis det var korrekt, ville det ikke være nødvendigt at lave statistik, fordi man altid ville se præcis det samme resultat. Haldane og Anscombe foreslog at lave en justering af antallene i en 2x2 tabel ved at lægge 0.5 til alle celler, så man undgår at antage, at en hændelse aldrig indtræffer. Udregn *OR* efter Haldane-Anscombes justering. I hvilken retning peger resultatet?

	Ændring i *RB1*	Ingen ændring i *RB1*
Ændring i *CDKN2A*	1	8
Ingen ændring i *CDKN2A*	0	2

Matematikøvelse 2. Hjernekræft: Betydning af flere mutationer - en statistisk undersøgelse

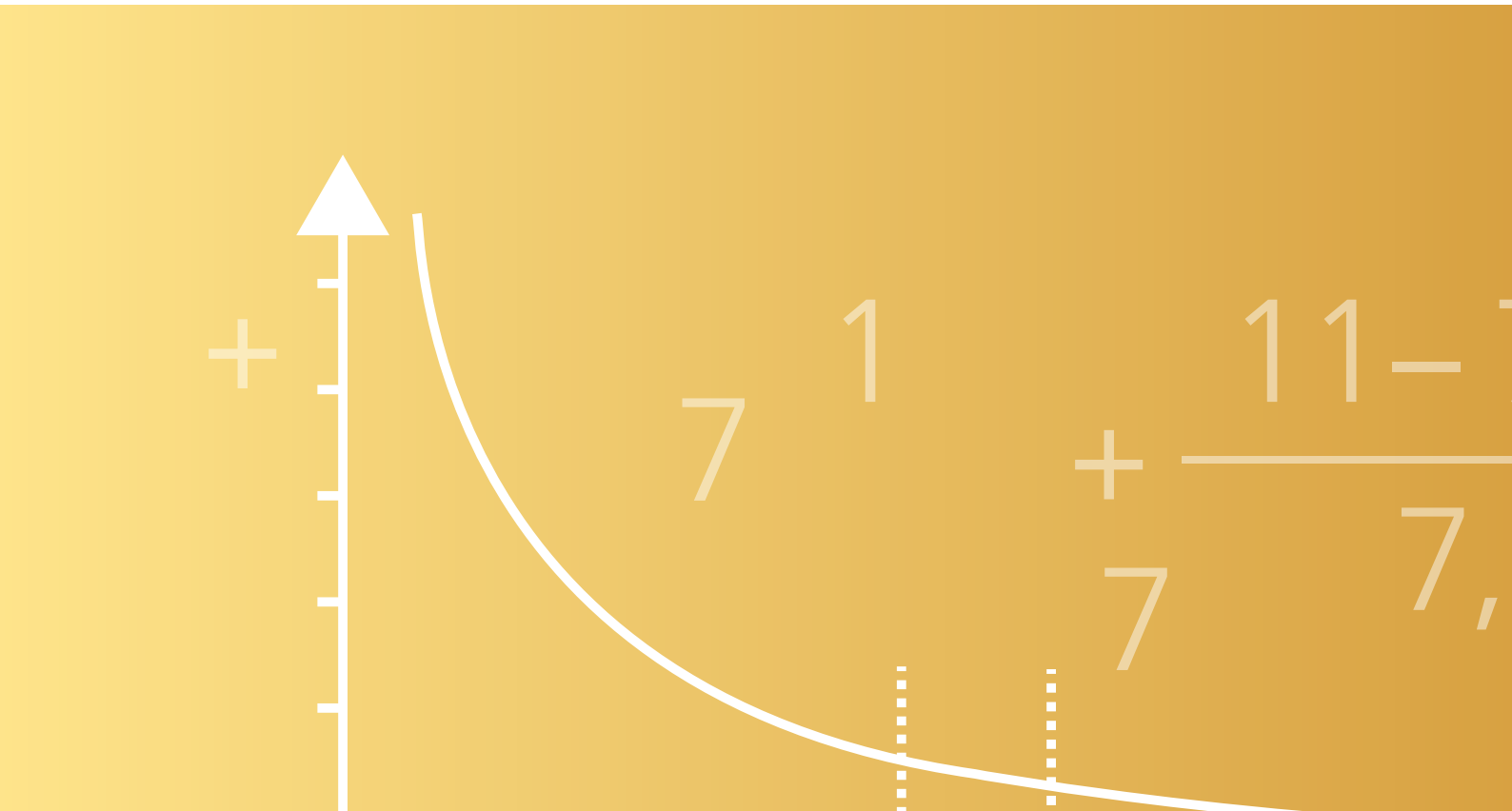
Introduktion

Vi tager udgangspunkt i data fra cBioPortal som ser nogenlunde således ud:



Vi skal dog bruge samme datasæt, som I bruger i Biotekøvelsen "*Hjernekræft: Mutationer i gener i den samme signaleringskaskade*". Vi skal ud fra data undersøge, om der er uafhængighed mellem forekomsten af gener hos patienter med denne kræfttype (hjernekræft) vha. et uafhængighedstest.

I denne forbindelse bestemmes en teststørrelse som er en sum af flere bidrag (hvilket vi skal vende tilbage til). Undersøgelse af disse bidrags (forholdsmæssige) størrelse kan desuden anvendes til at vurdere, om generne ofte forekommer sammen (co-occuring) eller sjældent sammen (mutually exclusive). Man kan også inddrage odds-ratio (OR) for at opstille en test der kan afgøre, om der er tale om gener der er "mutually exclusive" eller "co-occuring".



Den statistiske behandling - det skal I gøre og svare på

Vi tager som sagt udgangspunkt i data, som I normalt selv skal hente fra cBioPortal (punkt 1-3). Skal man have hjælp til dette, kan man følge vejledningen i bilag 1.

Alternativt kan man anvende Excel-arket Data_til_krydstabel.xlsx, hvor data er hentet og ordnet (husk at gemme filen på din egen computer). For at svare på punkt 4-8 kan man følge gennemregningen af eksemplet som er i bilag 2.

Udfør nu følgende:

1. Hent "de rigtige" data i cBioPortal.
2. Konverter og indhent data i Excel.
3. Find for hver patient i datasættet ud af, om der er ændringer i *CDKN2A, CDK4* og/eller *RB1*.
4. Lav krydstabeller for parrene *CDKN2A/CDK4, CDKN2A/RB1* og *CDK4/RB1*.
5. Opstil en hypotese for hver af parrene til at afgøre, om de er uafhængige.
6. Bestem i hvert tilfælde de forventede værdier.
7. Afgør på et 5% signifikansniveau, om hypoteserne kan forkastes.
8. Se på de enkelte bidrag og kommenter sammenhænge mellem ændringer i gener (måske co-occuring eller mutually exclusive). Udregn evt. OR, lav en hypotese og test denne.
9. Vend tilbage til biotekøvelsen "*Hjernekræft: Mutationer i gener i den samme signaleringskaskade*"

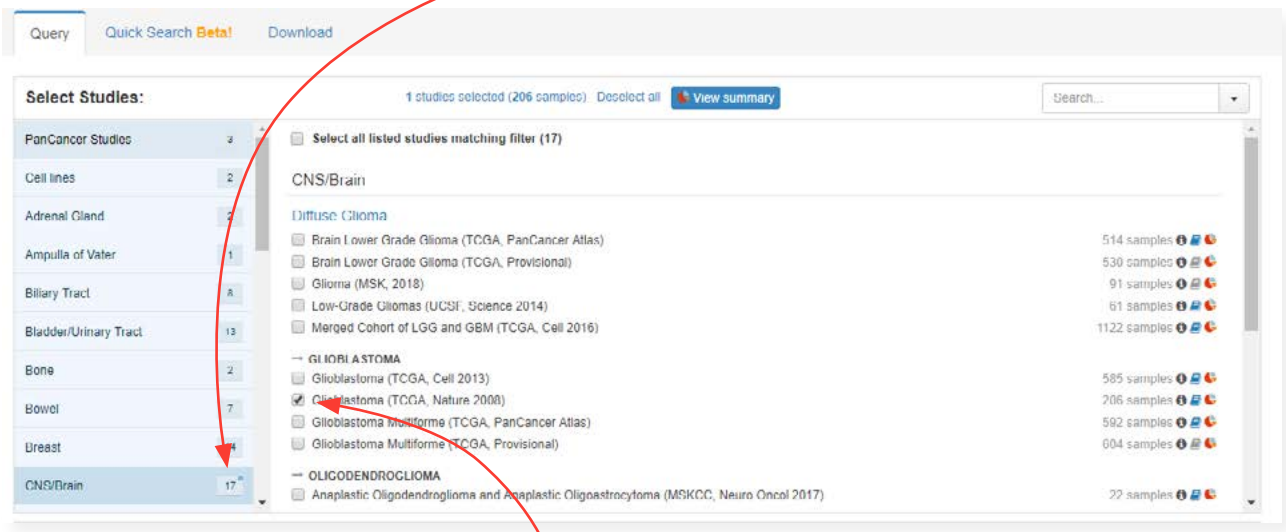
Bilag 1: Hente data fra cBioPortal og ordne dem i et Excel-regneark

Fremskaffe data

Ønsker man selv at hente data skal man:

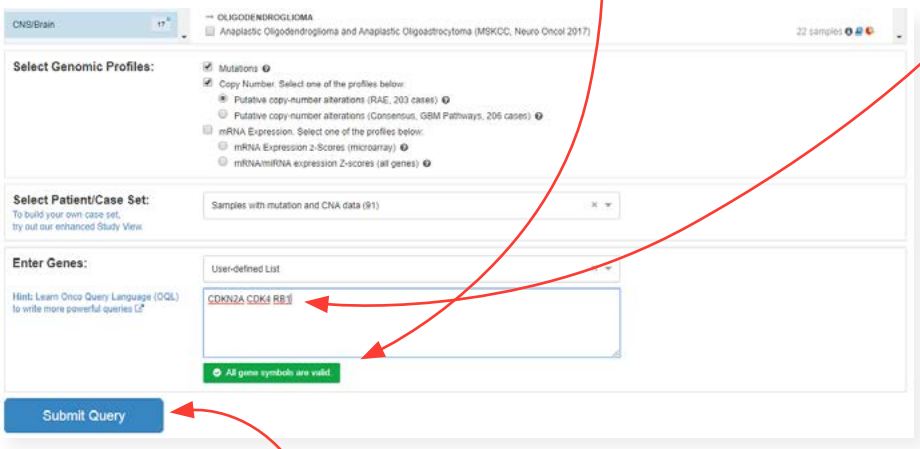
A. Gå til http://www.cbioportal.org/

B. Under "Select Studies" vælges "CNS/Brain"



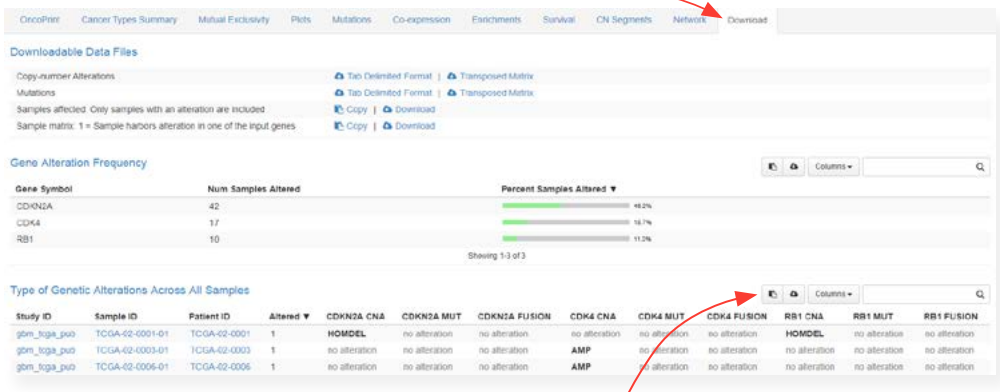
C. Så vælges "Glioblastoma (TCGA, Nature 2008)" og der trykkes på "Query by Gene.

D. Der scrolles længere ned på siden, og under "Enter genes" tastes *CDKN2A CDK4 RB1*, og det kontrolleres at "All gene symbols are valid".



E. Tryk på "Submit Query"

F. Efter lidt tid kommer der et resultatvindue, hvor fanen "Oncoprint" er valgt. Gå til fanebladet "Download"



G. Data kan nu hentes ved at downloade eller kopiere data. Data kan så senere indhentes eller indsættes i et regneark (f.eks. Excel)

Transformation af data

Data hentet fra CBioPortal har nedenstående struktur (der er tale om et udsnit), når de hentes ind i et regneark (Der er byttet om på rækker og søjler i forhold til den hentede fil)

Genes	Patient ID	TCGA-06-0221	TCGA-02-0001	TCGA-06-0214	TCGA-02-0083	TCGA-02-0010	TCGA-02-0046	TCGA-02-0060	TCGA-06-0210	TCGA-06-0211	TCGA-06-0214	TCGA-02-0083
CDKN2A		Deep	Deep	Deep	Trun...	Miss...			Deep	Deep	Deep	Trun...
CDKN2A	CNA	Deep	Deep	Deep	No alteration	No alteration	No alteration	No alteration	Deep	Deep	Deep	No alteration
CDKN2A	MUTATIONS	No alteration	No alteration	No alteration	Trun...	Miss...	No alteration	No alteration	No alteration	No alteration	No alteration	Trun...
CDKN2A	FUSION	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration
CDK4		Amp					Amp	Amp				
CDK4	CNA	Amp	No alteration	No alteration	No alteration	No alteration	Amp	Amp	No alteration	No alteration	No alteration	No alteration
CDK4	MUTATIONS	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration
CDK4	FUSION	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration
RB1	CNA		Deep									
RB1	CNA	No alteration	Deep	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration
RB1	MUTATIONS	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration
RB1	FUSION	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration	No alteration

Denne tabel omsættes til den noget kortere tabel som kun indeholder information om, hvorvidt der er en ændring på genet:

Gen	Ændring	TCGA-06-0221	TCGA-02-0001	TCGA-06-0214	TCGA-02-0083	TCGA-02-0010	TCGA-02-0046	TCGA-02-0060	TCGA-06-0210	TCGA-06-0211	TCGA-06-0214	TCGA-02-0083
CDKN2A	CNA/MUT/FUS	x	x	x	x	x			x	x	x	x
CDK4	CNA/MUT/FUS	x					x	x				
RB1	CNA/MUT/FUS		x									

Bilag 2: Opstilling af krydstabeller og hypotesetest for et eksempel

Opstilling af krydstabeller

På baggrund af tabellen som er illustreret nedenfor (som repræsenterer et udsnit af data), og som man har i f.eks. Excel opstilles parvise krydstabeller.

Gen	Ændring	TCGA-06-0221	TCGA-02-0001	TCGA-06-0214	TCGA-02-0083	TCGA-02-0010	TCGA-02-0046	TCGA-02-0060	TCGA-06-0210	TCGA-06-0211	TCGA-06-0214	TCGA-02-0083
CDKN2A	CNA/MUT/FUS	x	x	x	x	x			x	X	x	x
CDK4	CNA/MUT/FUS	x					x	x				
RB1	CNA/MUT/FUS		x									

Vi vil undersøge generne to og to. Med tre forskellige gener giver dette de tre forskellige par af gener (*CDKN2A*/*CDK4*, *CDKN2A*/*RB1* og *CDK4*/*RB1*). Ser vi på generne af *CDKN2A* og *CDK4* (dvs. RB1 ikke er interessant og udelades), kan det af tabellen ovenfor som indeholder et udsnit af data ses, at der er en patient med ændringer i begge gener (patient TCGA-06-0221), otte patienter der *kun* har ændring i genet *CDKN2A*, og to patienter der *kun har* ændring i genet *CDK4*. Der er ingen patienter i udsnittet som ikke har ændringer i mindst et af generne. Denne fremgangsmåde er anvendt på et langt større udsnit af data (50 patienter). Dette giver krydstabellen nedenfor til venstre. Ved tilsvarende argumentation for de andre par af gener (*CDKN2A*/*RB1* og *CDK4*/*RB1*) kan man lave de to andre krydstabeller (på de samme 50 patienter).

Krydstabel mellem *CDKN2A* og *CDK4*		Ændring i *CDK4*	
		Ja	Nej
Ændring i *CDKN2A*	Ja	3	21
	Nej	11	15

Krydstabel mellem *CDKN2A* og *RB1*		Ændring i *RB1*	
		Ja	Nej
Ændring i *CDKN2A*	Ja	1	23
	Nej	7	19

Krydstabel mellem *CDK4* og *RB1*		Ændring i *RB1*	
		Ja	Nej
Ændring i *CDK4*	Ja	0	14
	Nej	8	28

Hypotesetest - Et eksempel på behandling af data.

Data er nu placeret i en krydstabel. Vi vil nu tage udgangspunkt i eksemplet mellem *CDKN2A* og *CDK4*.

Vi laver nulhypotesen:

H_0: Tilstedeværelsen af en ændring i genet CDKN2A og en ændring i genet CDK4 er uafhængige.

Under denne hypotese er sandsynligheden for en ændring i genet *CDKN2A* ikke afhængigt af, om der er en ændring i genet *CDK4* eller ej. Da der i tabellen er 24 (3+21) patienter der har ændringen i genet *CDKN2A* ud af i alt 50 patienter, må sandsynligheden for at have ændringer i genet *CDKN2A*, p_{CDK2NA}, være

$$p_{CDKN2A} = \frac{24}{50}$$

For en tilsvarende argumentation omkring ændringer i genet *CDK4* må sandsynligheden for at have ændringer i genet *CDK4*, p_{CDK4}, være

$$p_{CDK4} = \frac{14}{50}$$

Dvs. sandsynligheden for en ændring i henholdsvis *CDKN2A* og *CDK4* under hypotesen er $p_{CDKN2A} = \frac{24}{50}$ og $p_{CDK4} = \frac{14}{50}$, mens sandsynlighederne for ikke at have ændring i generne er henholdsvis $\frac{26}{50}$ og $\frac{36}{50}$.

Vi kan nu bestemme de forventede værdier under nulhypotesen, hvilket gøres ved at bestemme sandsynlighederne og gange med det samlede antal patienter.

Sandsynligheden for at en patient har ændringer på genet eller ej i en bestemt kombination bestemmes som et produkt af sandsynligheder, da forekomsten af ændringer i *CDKN2A* og *CDK4* er uafhængige (under hypotesen).

Ændring i både *CDKN2A* og *CDK4* har sandsynligheden: $p_{CDKN2A} \cdot p_{CDK4} = \frac{24}{50} \cdot \frac{14}{50} = \frac{336}{2500}$

Ændring i *CDKN2A*, men ikke *CDK4* har sandsynligheden: $p_{CDKN2A} \cdot (1 - p_{CDK4}) = \frac{24}{50} \cdot \frac{36}{50} = \frac{864}{2500}$

Ændring i *CDK4*, men ikke *CDKN2A* har sandsynligheden: $(1 - p_{CDKN2A}) \cdot p_{CDK4} = \frac{26}{50} \cdot \frac{14}{50} = \frac{364}{2500}$

Ingen ændring i *CDKN2A* og i *CDK4* har sandsynligheden: $(1 - p_{CDKN2A}) \cdot (1 - p_{CDK4}) = \frac{26}{50} \cdot \frac{36}{50} = \frac{936}{2500}$

De forventede værdier kan nu bestemmes ved at gange med antallet, hvilket er gjort i tabellen til højre, mens tabellen til venstre indeholder de observerede værdier.

Observeret Krydstabel mellem *CDKN2A* og *CDK4*		Ændring i *CDK4*		
		Ja	Nej	Sum
Ændring i *CDKN2A*	Ja	3	21	24
	Nej	11	15	26
	Sum	14	36	50

Forventet Krydstabel mellem *CDKN2A* og *CDK4*		Ændring i *CDK4*		
		Ja	Nej	Sum
Ændring i *CDKN2A*	Ja	$\frac{336}{2500} \cdot 50 =$ 6,72	$\frac{864}{2500} \cdot 50 =$ 17,28	24
	Nej	$\frac{364}{2500} \cdot 50 =$ 7,28	$\frac{936}{2500} \cdot 50 =$ 18,72	26
	Sum	14	36	50

Bemærk, at hvis en eller flere af de forventede værdier er under 5, vil det ofte være ret problematisk, idet teststørrelsen typisk bliver for stor.

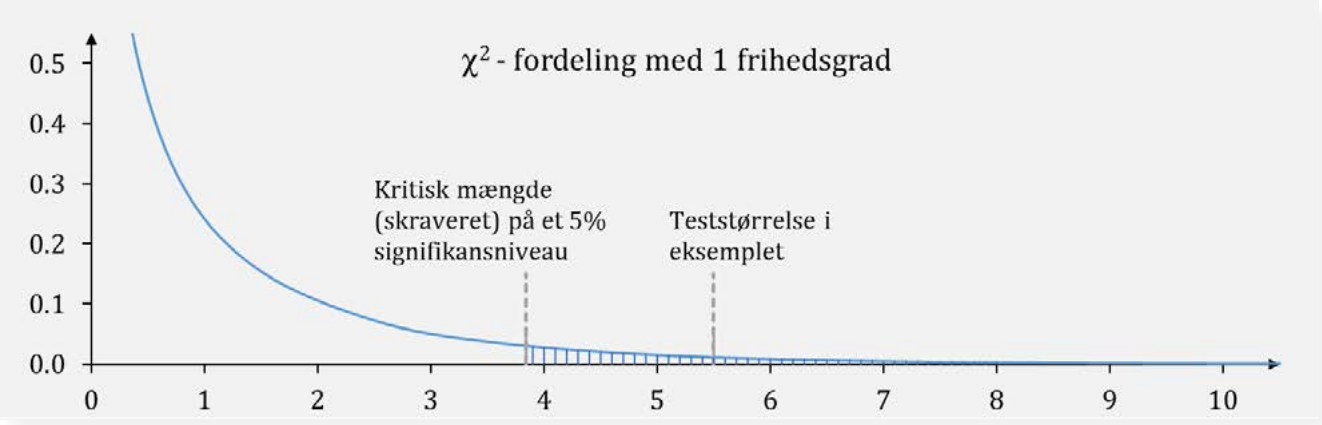
Teststørrelsen kan nu bestemmes ud fra de forventede og observerede værdier på følgende måde, da der er 4 forventede og 4 observerede værdier.

$$X^2 = \frac{(\text{obs}_1 - \text{forv}_1)^2}{\text{forv}_1} + \frac{(\text{obs}_2 - \text{forv}_2)^2}{\text{forv}_2} + \frac{(\text{obs}_3 - \text{forv}_3)^2}{\text{forv}_3} + \frac{(\text{obs}_4 - \text{forv}_4)^2}{\text{forv}_4}$$

Ved indsættelse af værdierne får vi:

$$X^2 = \frac{(3-6{,}72)^2}{6{,}72} + \frac{(21-17{,}28)^2}{17{,}28} + \frac{(11-7{,}28)^2}{7{,}28} + \frac{(15-18{,}72)^2}{18{,}72} \approx 2{,}059 + 0{,}801 + 1{,}901 + 0{,}739 = 5{,}500$$

Da der er tale om en 2×2 tabel, kan det afgøres, om vi skal forkaste hypotesen ved at sammenligne med en χ^2-fordeling med 1 frihedsgrad (antallet af frihedsgrader findes som (m – 1) · (n – 1) for en krydstabel med m søjler n og rækker). Ved denne sammenligning kan man også finde p-værdien. Sammenligningen sker typisk i jeres regneprogram, men her er det vist hvad der skal gøres



Da teststørrelsen ligger i det kritiske område (på et 5% signifikansniveau), skal vi forkaste hypotesen. Dvs. at en ændring i genet *CDKN2A* og en ændring i genet *CDK4 ikke* er uafhængige.

Det fremgår også, at *p*-værdien er ret lav (den kan beregnes til 0,019), hvilket betyder, at det er ret usandsynligt, at man ved en tilfældighed får så stor en teststørrelse, hvis hypotesen er sand. Faktisk er

Bidrag til teststørrelsen		Ændring i *CDK4*	
Krydstabel mellem *CDKN2A* og *CDK4*		Ja	Nej
Ændring i *CDKN2A*	Ja	2,059	0,801
	Nej	1,901	0,739

Forskel mellem observeret og forventet værdi		Ændring i *CDK4*	
Krydstabel mellem *CDKN2A* og *CDK4*		Ja	Nej
Ændring i *CDKN2A*	Ja	-3,72	3,72
	Nej	3,72	-3,72

der kun er 1,9% sandsynlighed for at få et værre resultat, hvis hypotesen er sand. Tilsvarende vil vi derfor også forkaste nulhypotesen på et 5% signifikansniveau.

Da vi nu ved, at vi kan forkaste hypotesen, kan det være gavnligt at se på bidrag til teststørrelsen og på forskellen mellem observerede og forventede værdier:

Af bidragene kan man typisk se, hvor "problemet" er ved at finde (usædvanligt) høje værdier.

Ser vi på forskellen mellem observerede og forventede værdier (observeret fratrukket forventet), ses det, at de felter, hvor der ændring i nøjagtigt et gen, er positive, svarende til at genet er observeret oftere end forventet under hypotesen. Dette tyder på, at ændringer i generne ofte forekommer alene (mutually exclusive).

Bemærkning: Havde feltet hvor der er ændringer i begge gener samtidigt været observeret hyppigere, end hvad forventes under hypotesen, kunne noget tyde på, at ændringer i generne ofte forekommer samtidigt (co-occuring).

Hvis man ikke vil nøjes med ovenstående overvejelser, kan man yderligere teste for mutually exclusive) og co-occuring gener ved at inddrage OR (odds-ratio) og test på baggrund heraf. Se evt., artiklen for flere detaljer.

Overlevelsesanalyse:
Kaplan-Meier kurver og log-rank test

Anne Lyngholm, Bjarke Hansen og Claus Ekstrøm | November 2019

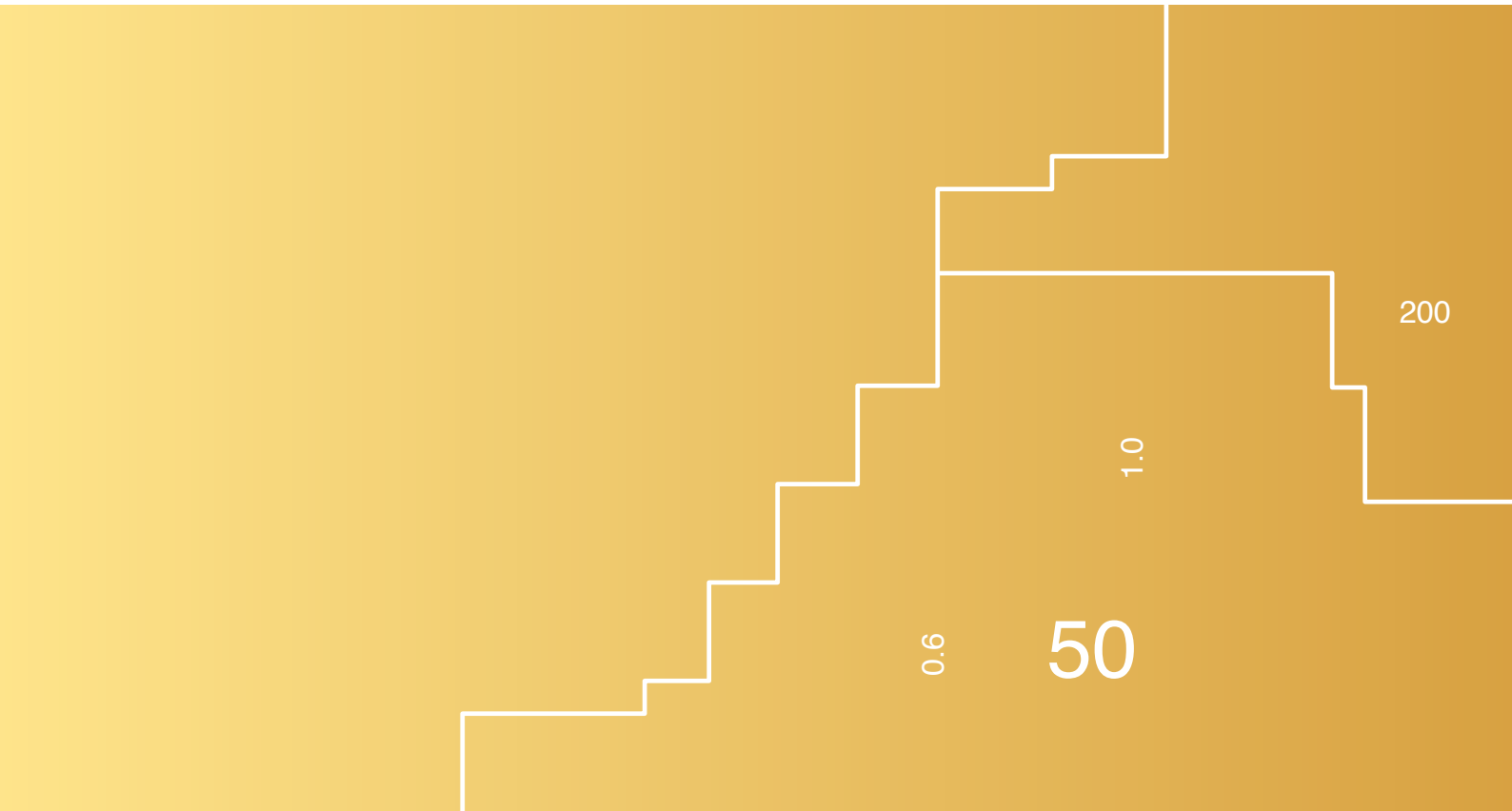
Det skal vi lære

Når I har været gennem materialet, skal I kunne svare på følgende:

- Hvad er censurering, og hvorfor er det vigtigt at tage højde for det?
- Hvordan beregner man en overlevelsessandsynlighed?
- Hvad viser en Kaplan-Meier kurve?
- Hvad er medianlevetiden?
- Hvad kan man konkludere ud fra et log-rank test?

Introduktion

Overlevelsesanalyse er en specifik gren af statistikken som handler om at estimere *tiden* til en bestemt hændelse f.eks. død eller tilbagevendende kræft. Vi fokuserer på hændelsen død og kalder i denne forbindelse tiden fra starttidspunket (f.eks. fødsel eller diagnose) til dødstidspunktet for overlevelsestiden. Med overlevelsesanalyse er man ofte interesseret i at undersøge mulige forskelle i fordelingen af overlevelsestid mellem grupper (f.eks. behandlingsformer). En metode til at visualisere fordelingen af overlevelsestider er Kaplan-Meier kurver. Sammenligning af overlevelsestider mellem forskellige grupper kan ske vha. af Kaplan-Meier kurver; men for at afgøre om der er forskel i overlevelsestiden har man brug for et såkaldt log-rank test.



Eksempel: Effekten af *TP53* mutation på overlevelse

Dette er et eksempel på anvendelse af Kaplan-Meier kurver og log-rank test til at undersøge effekten af en mutation af *TP53*. 42 patienter - 21 med og 21 uden en mutation af *TP53* - var med i forsøget. Formålet med forsøget var at undersøge, om patienter med mutationen levede kortere end de patienter uden mutationen - altså om der er forskel i overlevelsestiden mellem de to grupper. Studiet startede den 1. juni 1963 og varede i 36 måneder. Patienterne blev derfor fulgt i op til 36 måneder efter deres start af ovariekræftbehandling. figur 1 viser data noteret for 10 ud af de 42 patienter i forsøget.

	A	B	C	D	E	F
1	Patient_ID	Gruppe	Måned_Behandling	Måned_FollowUp	Overlevelsestid	Status
2	1	Mutation	01-04-1965	01-04-1966	12	Død
3	2	Mutation	01-02-1964	01-01-1965	17	Død
4	3	Ikke mutation	01-10-1965	01-06-1966	9	I live
5	4	Ikke mutation	01-11-1964	01-06-1966	19	I live
6	5	Mutation	01-06-1963	01-04-1965	22	Død
7	6	Ikke mutation	01-10-1963	01-04-1964	6	Død
8	7	Ikke mutation	01-05-1964	01-06-1966	25	I live
9	8	Ikke mutation	01-06-1965	01-04-1966	10	Død
10	9	Mutation	01-06-1964	01-08-1964	2	Død
11	10	Mutation	01-12-1963	01-12-1964	12	Død

Figur 1. Udsnit af data for de første ti personer i forsøget.

Fra dataudsnittet ses det, at patienterne startede behandlingen på forskellige tidspunkter, de har forskellige længde af opfølgning og nogle af patienterne døde under studieperioden mens andre var i live da forsøget sluttede. I overlevelsesanalyse er det meget almindeligt, at man ikke observerer alle patienternes dødstidspunkter, fordi forsøget stopper inden alle patienter er døde.

Overlevelsesanalyse

Til overlevelsesanalyser skal der for hver person bruges følgende data:

- **Overlevelsestiden**, som er tiden mellem starttidspunktet og sluttidspunktet. Sluttidspunktet er enten dødstidspunktet eller sidste tidspunkt, hvor det blev bekræftet, at patienten endnu ikke var død. Hvis patienten stadig er i live kaldes sluttidspunktet også censureringstidspunktet.
- **Status** ved slutningen af studiet. Der er to muligheder. Enten er patienten ''Død" eller også er patienten ''Censureret''/''I live".
- **Andre variable** f.eks. genmutation, behandling, alder, køn, osv.

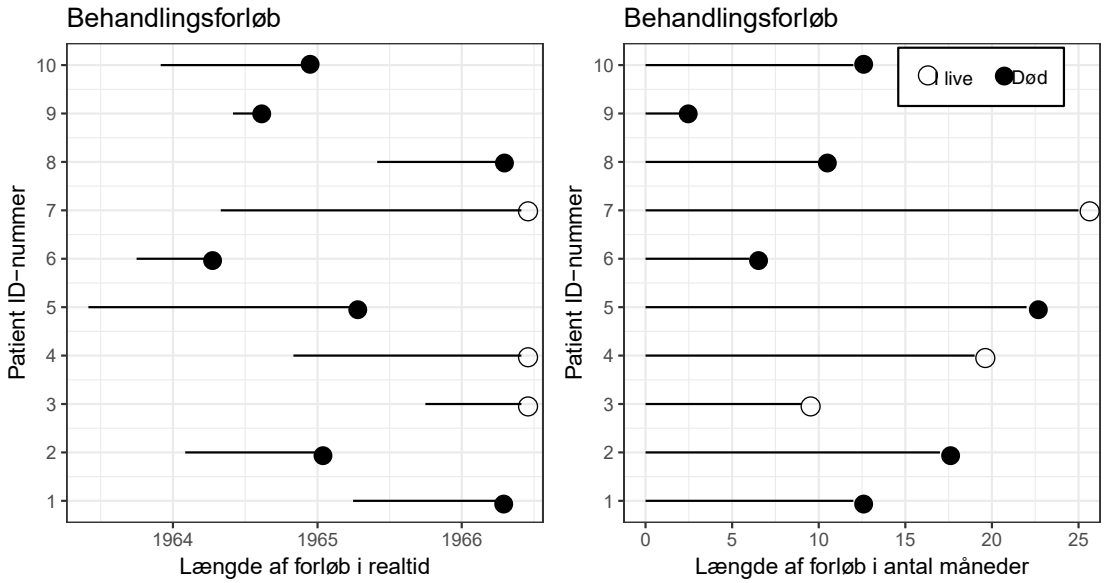
Alle tre typer indhold kan findes i figur 1: **Overlevelsestid** er tidsintervallet altså længden af overlevelse efter start af behandling, **Status** er om patienten var død eller i live ved sidste opfølgningstidspunkt, mens **Gruppe** er hvorvidt personen har mutationen eller ej.

Inden vi forsætter til analysen, er det nødvendigt at forklare begrebet censurering, og hvorfor det er vigtigt i denne type analyse.

Hvis en person er blevet censureret, så betyder det, at man ikke kender personens dødstidspunkt, men at man til gengæld ved, at personen i hvert fald var i live indtil censureringstidspunktet. Denne information er nyttig, for i overlevelsesanalyse sammenligner man antallet af personer under risiko for at dø i en bestemt tidsperiode med antallet af personer, som rent faktisk døde i samme tidsperiode.

Antallet af personer som er under risiko for at dø i en bestemt tidsperiode kaldes for risikosættet. Alle er med i risikosættet indtil de dør eller bliver censurerede (også dem der dør for man er nødt til at være i risiko for at dø for rent faktisk at kunne dø). På den måde bidrager de censurerede personer stadig til analysen, fordi vi ved, at de har været under risiko for at dø, uden at de faktisk gjorde det.

En grafisk måde at visualisere betydningen af censurering og overlevelsesanalyse er vist i figur 2. På plottene vises behandlingsforløbet for patienterne fra figur 1. Linjerne på grafen repræsenterer starten på behandlingen og slutningen af den enkelte patients forløb med enten censurering (tom cirkel) eller død (fyldt cirkel). Det venstre plot, viser hvordan man observerede patienterne i realtid, og det højre plot visualiserer de data, som man vil benytte i en overlevelsesanalyse, nemlig behandlingsforløbet i absoluttid fra behandlingsstart til død eller censurering (follow-up).



Figur 2. Plots over forløbet efter behandling i hhv. realtid og absolut tid (antal af måneder).
Censurerede observationer er markeret med en tom cirkel imens død er markeret med en fuld cirkel.

Kaplan-Meier overlevelseskurven

Data fra eksemplet med *TP53* er vist i tabellen nedenfor, hvor overlevelsestiden målt i måneder er opdelt efter, hvorvidt der er mutation på genet *TP53*. Tal markeret med '+' indikerer, at tidspunktet svarer til en censurering, så man ved bare, at den pågældende patient ikke var død inden overlevelsestiden.

Overlevelsestid i ikke-mutationsgruppen	Overlevelsestid i mutationsgruppen
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Fra tabellen kan vi i mutationsgruppen aflæse, at der var tre patienter som døde efter sjette måned og én patient som blev censureret ved sjette måned. Gennemsnittet af de 21 tal for ikke-mutationsgruppen

i tabellen er 17.1 måneder, men dette tal er ikke den reelle gennemsnitlige overlevelsestid for ikke-mutationsgruppen.

Hvis man fejlagtigt behandlede censureringstidspunkter som dødstidspunkter ville overlevelsestiden blive estimeret til at være mindre, end den skulle være, fordi værdierne markeret med '+' angiver den kortest mulige overlevelse for den enkelte censurering. For at få et bedre overblik over data kan vi lave følgende tabel, en såkaldt livstabel (enkelte værdier er udeladt).

Længde af overlevelse i rækkefølge i mdr	Antal døde	Antal censurerede	Antal i risikosættet ved periodens start	Overlevelses-sandsynlighed
0	0	0	21	1
6	3	1	21	$1 \cdot 18/21 = 0.8571$
7	1	1	17	$0.8571 \cdot 16/17 = 0.8067$
10	1	2	15	$0.8067 \cdot 14/15 = 0.7529$
13	1	0	12	$0.7529 \cdot 11/12 = 0.6902$
16				
22				
23				

For hver måned, hvor en patient dør, tæller vi antallet af patienter som døde og antallet af patienter som blev censureret. For hver periode er sandsynligheden for at overleve lig antallet af patienter som stadig er i live ved periodens slutning i forhold til dem som startede med at være i live ved periodens start.

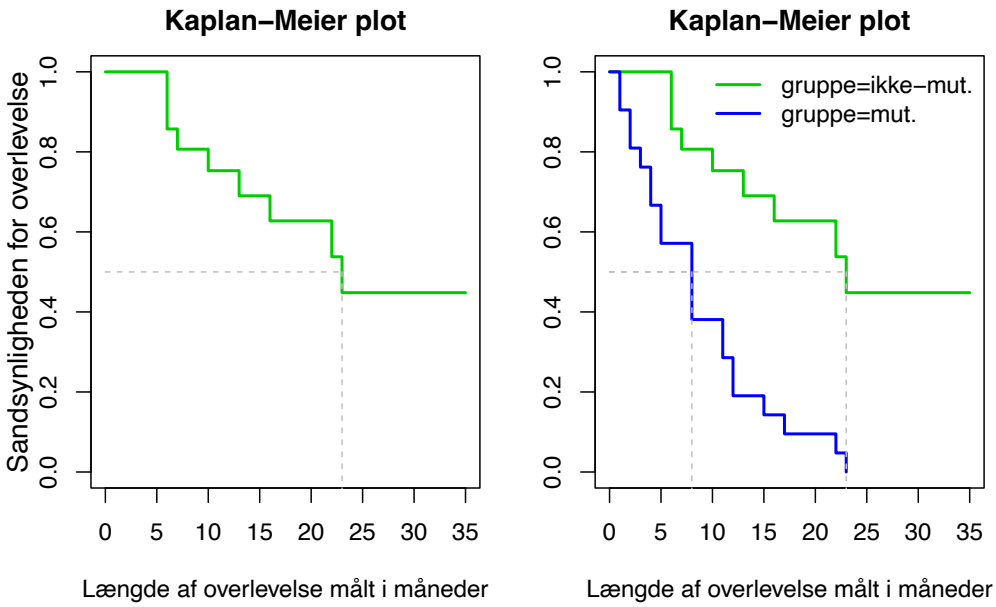
Lad os tage et eksempel fra tabellen: Op til seks måneder efter start af behandlingen var alle 21 patienter stadig med i forsøget og i live. Risikosættet op til 6. måned er derfor på 21 personer. I 6. måned dør tre patienter og sandsynligheden for at dø bliver da $\frac{3}{21}$ = 14.3%. Den tilsvarende sandsynlighed for at overleve den 6. måned er derfor $\frac{18}{21}$ = 1– $\frac{3}{21}$ = 85.7% . I beregningen indgår de censurerede patienter ikke, men de fjernes i stedet fra risikosættet *efter* censureringstidspunktet. Patienter er med i risikogruppen op til censureringstidspunktet, da de i denne periode på lige fod med andre er i risiko for at dø.

Overlevelsessandsynligheden beregnes ud fra følgende logik: Til tid 0 er alle patienter stadig i live og overlevelsessandsynligheden må derfor være 100% (skrevet som 1 i livstabellen). Op til den sjette måned er 18 ud af 21 patienter stadig i live, så overlevelsessandsynligheden er 85.71%. Da tre patienter døde og en patient udgik af studiet grundet censurering reduceres risikogruppen efter sjette måned til 17 patienter. Ud af de 17 patienter, som er tilbage i studiet efter sjette måned, døde en patient i den syvende måned. Det betyder, at sandsynligheden for at dø i den syvende måned, **hvis man allerede havde overlevet den sjette måned** er 1/17 = 5.9% og sandsynligheden for at overleve er 16/17 = 94.1%. Selvom denne sandsynlighed er brugbar, er vi mere interesseret i overlevelsessandsynligheden fra start og helt til den syvende måned. Denne sandsynlighed er produktet mellem sandsynligheden for at overleve op til sjette måned og den betingende sandsynlighed for at overleve op til syvende måned så 0.8571 · 16/17 = 0.8067 = 80.67%. For at overleve til måned 7 skal man altså først overleve til måned 6, og derefter skal man yderligere overleve til måned 7.

Du kan nu regne øvelse 1a

Overlevelsessandsynligheden, beregnet i livstabellen, kan visualiseres over tid ved brug af et Kaplan-Meier plot (figur 3). På venstre plot er ikke-mutationsgruppens overlevelsessandsynlighed plottet mod længden af behandlingsforløbet målt i måneder.

Kaplan-Meier kurven anvender de censurerede tidspunkter og kurven er derfor baseret på alle data til rådighed. Vi kan på kurven observere sandsynligheden for overlevelse til forskellige tidspunkter. Vi kan på det venstre plot se, at der er 50 procent sandsynlighed for at overleve op til 23. måned. På det højre plot ses, at for gruppen med mutationer er sandsynligheden for at overleve 8 måneder helt nede på 50%.



Figur 3. Venstre plot: Kaplan-Meier plot for ikke-mutationsgruppen af patienter. **Højre plot:** Kaplan-Meier plot over både mutation- og ikke-mutationsgruppen.

På figur 3 ses en markant forskel på de to kurver, og gruppen uden mutationer har betydelig bedre overlevelsestid end gruppen med mutationer. Man er fristet til at konkludere, at personer uden mutationen lever længere end dem med mutationen i *TP53*, men for at kunne lave sådan en konklusion skal vi tage overlevelseskurvernes usikkerhed i betragtning. For at afgøre om der er en statistisk forskel på overlevelsesfunktionerne kan man teste hypotesen om, hvorvidt de kunne være sammenfaldende. Dette gøres med et såkaldt log-rank test.

Du kan nu regne øvelse 3

log-rank testet

Med et log-rank test kan vi sammenligne to Kaplan-Meier kurver og undersøge, om de overordnet er ens. Selve formlen for log-rank testet er simpel, men udregningen er ret besværlig. Vi vil derfor kun præsentere dele af testet her.

Hypotesen, som testes med et log-rank test, er, at de to kurver stammer fra samme overlevelsesfunktion. Hvis de to kurver stammer fra grupper med samme overlevelsesfunktion, så forventer vi, at begge de observerede overlevelsessandsynligheder ligner de forventede overlevelsessandsynligheder fra den fælles underliggende kurve, som vi påstår de kommer fra. I praksis undersøges dette ved at sammenligne det observerede antal døde med de forventede antal døde (under hypotesen) ved hvert tidspunkt hvor en patient dør. Tabellen over kræftpatienter kan her bruges som et eksempel:

	Antal døde		Riskosæt		Forventet døde		Obs.-forventet	
Tidspunkt	Ikke-mutation	Mutation	Ikke-mutation	Mutation	Ikke-mutation	Mutation	Ikke-mutation	Mutation
1	0	2	21	21	1	1	-1.00	1.00
2	0	2	21	19	1.05	0.95	-1.05	1.05
3	0	1	21	17	0.55	0.45	-0.55	0.55
…	…	…	…	…	…	…	…	…
6	3	0	21	12	1.91	1.09	1.09	-1.09
…	…	…	…	…	…	…	…	…

Fra tabellen ses, at efter 1. måned i behandling er to patienter fra mutationsgruppen døde mens ingen patienter i gruppen uden mutationer er døde. Hvis de to grupper skal have ens Kaplan-Meier kurver (som er nulhypotesen), så skal overlevelsessandsynligheden være ens i de to grupper. Når to ud af 42 patienter dør er overlevelsessandsynligheden til 1. måned efter behandlingsstart 2/42 = 4.8%, det forventede antal af døde i hhv. mutation- og ikke-mutationsgruppen er $0.048 \cdot 21 = 1$ patient.

Der er i alt 17 forskellige tidspunkter, hvor patienter dør, og beregningerne til log-rank statistikken minder om chi-i-anden statistikken, fordi man i begge tilfælde bruger det totale observerede og forventende antal per gruppe. Det observerede antal døde i ikke-mutationsgruppen er 9 patienter, mens det er 21 i mutationsgruppen. Summerer vi antallet af forventede døde i ikke-mutationsgruppen får vi 19.26 patienter, mens det for mutationsgruppen giver 10.74 døde. Det vil sige, at der er en forskel på $9-19.26=-10.26$ patienter i ikke-mutationsgruppen og en på $21-10.74=10.26$ patienter i mutationsgruppen. Selve beregningen for den approksimative log-rank statistik giver:

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - F_i)^2}{F_i} = \frac{(9-19.26)^2}{19.26} + \frac{(21-10.74)^2}{10.74} = 15.793$$

Her står O_i for Observeret, F_i for Forventet og K for antallet af Kategorier (hvilket svarer til antallet af grupper, der sammenlignes) og $\sum_{i=1}^{K}$ betyder, at vi summerer over alle de grupper, som vi har til rådighed (i vores eksempel de to grupper mutation og ikke-mutation). Den kritiske værdi for en chi-i-anden fordeling med en frihedsgrad er på et 5% signifikansniveau 3.84. Log-rank statistikken er i eksemplet bestemt til 15.793 altså langt over 3.84 og den tilhørende *p*-værdi er under 0.0001. Vi afviser derfor hypotesen om, at Kaplan-Meier kurverne stammer fra den samme overlevelsesfordeling. Kurverne er altså forskellige, hvilket også blev antydet i plottet.

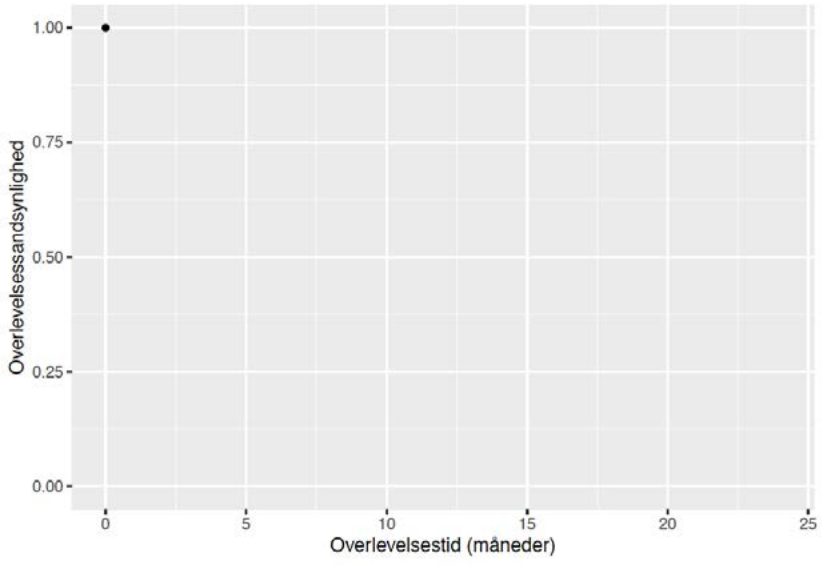
Bemærk, at log-rank testet ikke fortæller os noget om størrelsen på forskellen i overlevelse mellem grupperne. log-rank testet fortæller kun, *om* der er forskel, men ikke hvor stor forskellen er, eller hvilken gruppe, der klarer sig bedst. For at kunne svare på sidstnævnte må man se på Kaplan-Meier plottet, og se, hvordan de to kurver ligger i forhold til hinanden.

Hvad skal være opfyldt for at man må lave Kaplan-Meier plots og log-rank tests?

Der er få men vigtige antagelser for at bruge Kaplan-Meier kurver og log-rank testet. For det første skal patienterne være uafhængige af hinanden. Desuden er det vigtigt, at censureringen ikke skyldes den prognose, som patienten har. Det vil sige, at hvis patienten censureres, fordi de bliver sat i anden akutbehandling er censureringen ikke uafhængig af patients lavere overlevelsessandsynlighed. Ligeledes hvis patienten får det meget bedre og derfor flytter og bliver censureret, så er censurering ikke uafhængig, men skyldes at patientens overlevelsessandsynlighed er "god".

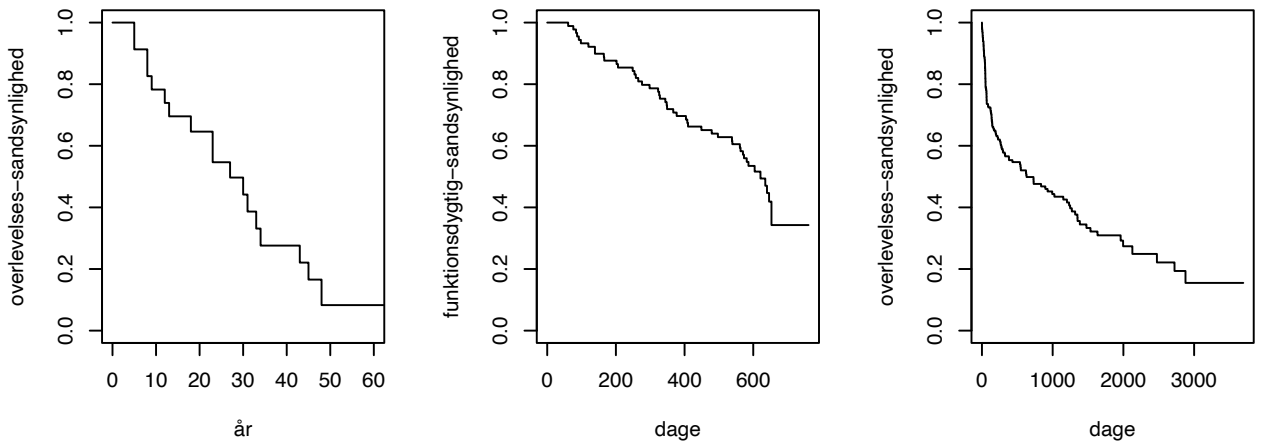
Øvelser

1. **Beregning af overlevelsessandsynlighed**
   a. Udfyld den resterende del af livstabellen (s.110) og lav en tilsvarende for mutationsgruppen.
   b. Skitsér derefter selv Kaplan-Meier plot ud fra tabellerne. Du kan bruge figur 4 som skabelon.



Figur 4. Skitse til at besvare opgave 1

2. **Median overlevelsestid**. Hvad er (cirka) median overlevelsestiden for hver af de tre Kaplan-Meier plots vist på figur 5?



Figur 5. Tre plots med hver deres hændelse.

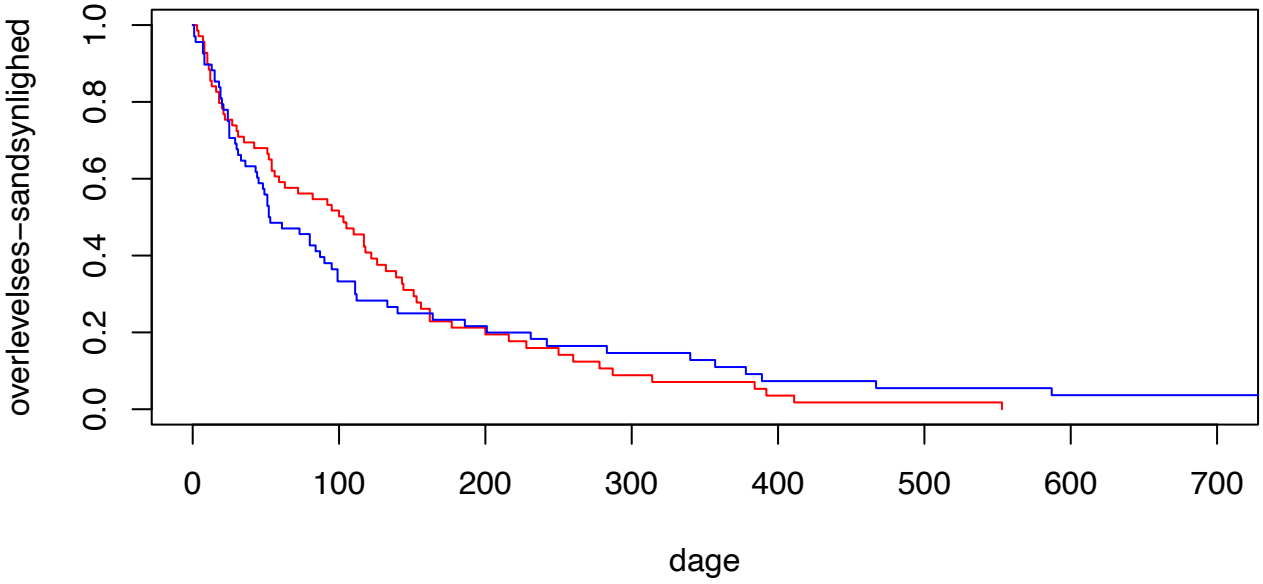
3. **Bedste behandling afhængig af tid**. Alle tre spørgsmål skal besvares ud fra nedenstående Kaplan-Meier plot, hvor to overlevelsestider siden diagnose er visualiseret over tid (hhv. illustreret ved en rød og blå kurve) på figur 6.

   a. Hvilken gruppe har den bedste overlevelsesprognose indtil dag 160?

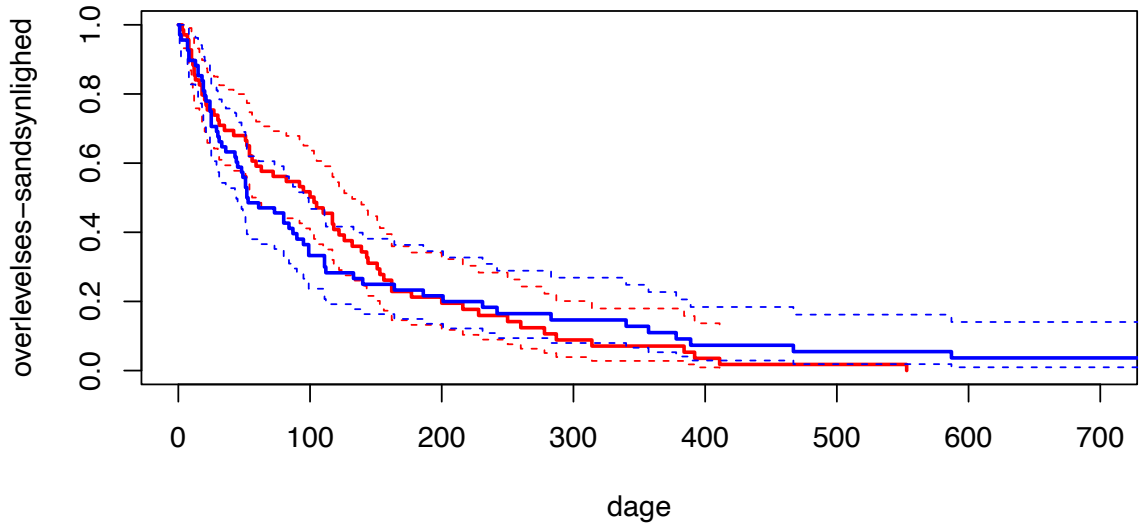
   b. Hvilken gruppe har den bedste overlevelsesprognose efter dag 160?

   c. Hvilken gruppe har den længste median overlevelsestid?

   Vi medtager nu usikkerheden på kurverne i form af 95% konfidensinterval. Dette er vist på figur 7.



Figur 6. Kaplan Meier kurver over to forskellige behandlingsgruppers overlevelsessandsynlighed afhængig af dage.



Figur 7. Kaplan-Meier kurver over to forskellige behandlingsgruppers overlevelsessandsynlighed afhængig af dage. Inklusiv 95% konfidensinterval.

d. Hvad er fortolkningen af de to 95% konfidensintervaller til dag 200? (brug din almindelige viden om konfidensitervaller til at svare på spørgsmålet).

e. Vil du konkludere på baggrund af konfidensintervallerne, at der er statistisk signifikant forskel på kurverne? Kommenter dit svar.

4. **log-rank og konklusion**. De følgende data er fra en undersøgelse, hvor man er interesseret i effekten af aktivering af et specifikt gen på overlevelsessandsynligheden. Husk, at tal markeret med **+** er censurerede data.

a. Hvilken hypotese kan vi teste med log-rank testet? Beskriv hypotesen i ord.

b. Udfyld den manglende information i tabellen nederst i denne øvelse.

Gruppe 1: Ikke aktiv	Gruppe 2: Aktiv
12.3+, 5.4, 8.2, 12.2+, 11.7, 10.0, 5.7, 9.8, 2.6, 11.0	5.7, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 11.4, 2.6, 9.8

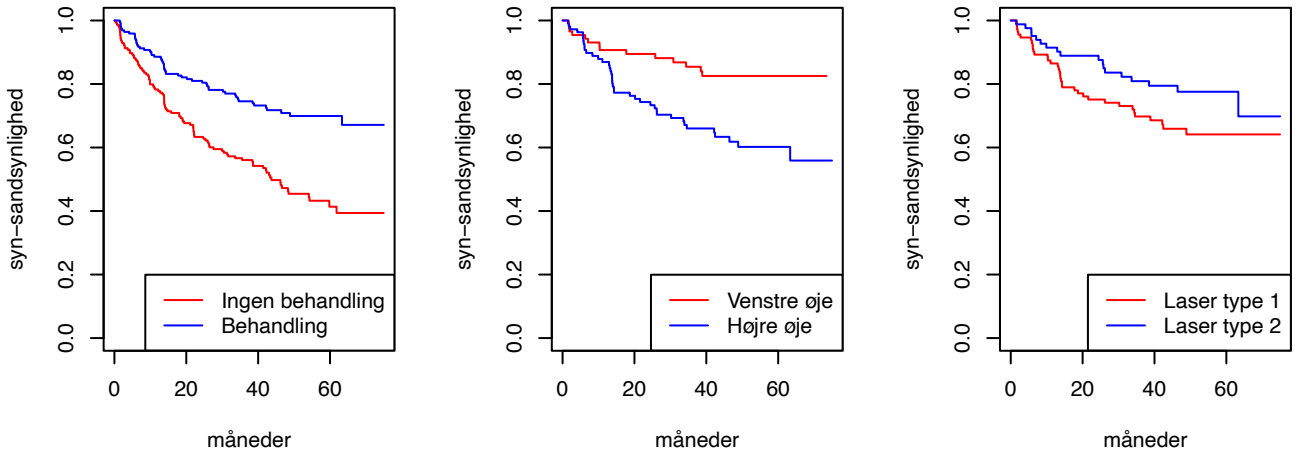
c. Summér for begge grupper henholdsvis antallet af både forventede og observerede døde ud fra den fyldte tabel.

d. Beregn log-rank testet ud fra de forventede og observerede døde, som du lige har beregnet.

e. Den kritiske værdi for log-rank testet for to grupper er 3.84. Hvis værdien for log-rank testet er over den kritiske værdi, så afviser vi hypotesen og *p*-værdien er tilsvarende under 0.05. Kommenter på dit resultat i ord, hvor du både kommer ind på hypotesen som er blevet testet, log-rank test værdien samt *p*-værdi (over 5% eller under 5%).

	Antal døde		Riskosæt		Forventet døde		Obs.-forventet	
Tidspunkt	Ikke-mutation	Mutation	Ikke-mutation	Mutation	Ikke-mutation	Mutation	Ikke-mutation	Mutation
2.6	1	1	10	10	1.00	1.00	0.00	0.00
2.9	0	1	9	9	0.50	0.50	-0.50	0.50
4.1	0	1	9	8	0.53	0.47	-0.53	0.53
4.2	0	1	9	7	0.56	0.44	-0.56	0.56
5.4	1	0	9	6	0.60	0.40	0.40	-0.40
5.7	1	1	8	6	1.14	0.86	-0.14	0.14
8.2	1	0	7	5	0.58	0.42	0.42	-0.42
8.3	0	1	6	5	0.55	0.45	-0.55	0.55
8.4	0	1	6	4	0.60	0.40	-0.60	0.60
9.1								
9.8								
10.0					0.83	0.17		
11.0	1	0						
11.4	0	1						
11.7	1	0			1.00	0.00	0.00	0.00

5. **Konklusioner**. Vi vil i denne øvelse se på et studie over forebyggelse af blindhed forårsaget af diabetes. Hændelsen vi undersøger i denne opgave er at blive blind. Figuren herunder viser 3 plots med Kaplan Meier kurver. Det første plot ser på effekten af behandling med laser på tid-til-blind (overlevelsestiden). Det næste plot ser på effekten af laser på hhv. højre og venstre øje. Det sidste plot ser på effekten af to slags laser-behandlingsmetoder.



Første studie har et log-rank test på 22.2, hvilket giver en *p*-værdi mindre end 0.001. Andet studie har et log-rank test på 9.5, hvilket giver en *p*-værdi på 0.002. Tredje studie har et log-rank test på 3, hvilket giver en *p*-værdi på 0.09.

a. Kommenter til at starte med kun på Kaplan-Meier kurverne. Hvad er din umiddelbare konklusion per studie? Hvilken effekt ser du per plot?

b. Inddrag nu log-rank testet og den tilhørende *p*-værdi. Stemmer disse overens med din konklusion per plot i det overstående spørgsmål?

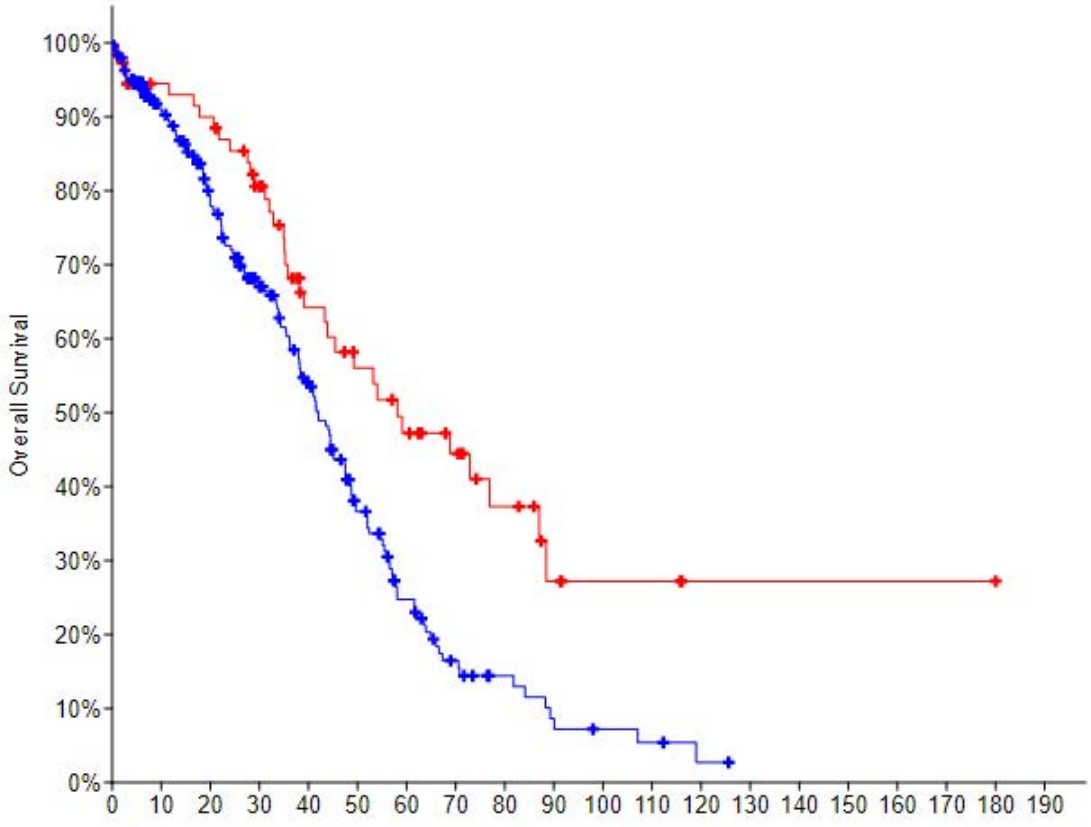
c. Formuler konklusionerne med først at beskrive hypotesen som testes og derefter resultaterne (Kaplan-Meier kurverne, log-rank testet og *p*-værdien).

Matematikøvelse 3. Æggestokkræft: Overlevelsesanalyse - en statistisk behandling

Intro

Vi tager udgangspunkt i diagrammet nedenfor, som I allerede har set i forbindelse med biotekøvelsen *"Æggestokkræft: Mutationer i BRCA1 og BRCA2 og deres betydning for overlevelse"*.

Diagrammet viser som bekendt overlevelseskurver for patienter med æggestokkræft. Den røde kurve viser patienter med ændringer i mindst et af generne *BRCA1* og *BRCA2*, og den blå viser patienter uden ændringer i *BRCA1* og *BRCA2*.



Figur 1: Overlevelseskurver for patienter med æggestokkræft. Den røde kurve er for patienter med ændringer i mindst et af generne *BRCA1* og *BRCA2* og den blå for patienter uden ændringer i *BRCA1* og *BRCA2*. Tiden 0 svarer til diagnosetidspunktet.

Vi vil i denne øvelse lave dele af kurverne manuelt for at blive lidt klogere på, hvad de egentlig viser. Efterfølgende vil vi se på det såkaldte log-rank test som kan bruges til at bestemme, om der er en signifikant forskel på de to overlevelseskurver.

Den statistiske behandling - eksempler og opgaver I skal svare på

I modsætning til de andre statistiske undersøgelser vil der i dette tilfælde være et overlap mellem eksempler, og det data som I arbejder på. Dette skyldes, at eksemplerne er på nøjagtig det data, som I også selv skal behandle.

Patientdata - en forklaring på transformation af data

Data til selv at lave overlevelseskurverne kan trækkes ud af cBioPortal. Det kræver dog et ret stort overblik af få fat i de rigtige data. Ønsker man selv at hente data, kan man se hvordan i bilag 1.

Vi vil i stedet her anvende den mere strukturerede Excel-fil "Data_til_KM_kurver.xlsx" som indeholder væsentlige data. Husk at gemme filen på din egen computer.

Udvalgte data fra syv patienter med æggestokkræft trukket ud fra cBioPortal kan ses i nedenstående tabel til venstre. De udvalgte data transformeres(oversættes) til en tabel som den nedenstående til højre. De transformerede findes i Excel-filen "Data_til_KM_kurver.xlsx" for langt flere patienter. Det er de transformerede data, der skal anvendes til at lave overlevelseskurverne.

Patient ID	Overall Survival (Months)	Overall Survival Status
TCGA-61-2109	20,66	DECEASED
TCGA-57-1584	21,12	LIVING
TCGA-24-0975	21,78	DECEASED
TCGA-57-1582	24,02	DECEASED
TCGA-23-1026	26,81	LIVING
TCGA-25-1625	27,60	DECEASED
TCGA-24-2035	28,16	DECEASED

Oversættelse →

Overlevelse (måneder)	Døde	Censurerede (inden næste dødstidspunkt)
20,66	1	1
21,78	1	
24,02	1	1
27,60	1	
28,16	1	

Den venstre tabel er transformeret til den højre påfølgende måde:

- Der laves et antal rækker, som svarer til døde patienter.
- De døde patienter sorteres efter overlevelsestid
- Det angives, hvor mange der døde til dette tidspunkt.
- Til slut angives, hvor mange der censureres, dvs. hvor mange der ikke med garanti overlever fra dette til næste dødstidspunkt.

Kaplan-Meier overlevelseskurver (KM-kurver) - Fremstilling

Excel-filen "Data_til_KM_kurver.xlsx" indeholder transformerede patientdata udgivet i Nature og svarer til dem, I ser på i biotekdelen.

Tabellerne indeholder data fra 73 patienter med ændringer på *BRCA1* og/eller *BRCA2* og desuden data fra 242 patienter med ændringer på andre gener. I ser i eksemplet undervejs kun på de 73 patienter med ændringer på *BRCA1* og/eller *BRCA2*, men data for de resterende patienter kan behandles på samme måde.

I Excel-arket findes allerede overlevelse, døde og censurerede og vi ønsker at finde sandsynligheden for overlevelse fra start til et bestemt tidspunkt. Hertil får vi undervejs brug for at bestemme det såkaldte risikosæt og sandsynligheden for at overleve fra et dødstidspunkt til et andet dødstidspunkt. Vi skal altså ende med en tabel som indeholder nedenstående information.

Overlevelsestid:	Angiver hvor længe der gik før død eller censur.
Døde	Angiver antal der dør efter at have overlevet dette tidsrum
Censurerede:	Angiver antallet af personer som har overlevet længere end overlevelsen for sidste overlevelsestid og højst denne overlevelsestid (Patienterne er typisk levende, men har kun haft diagnosen i kort tid).
Risikosæt:	Antallet af personer som kan dø i næste tidsrum (beregnes som risikosættet fra sidste tidsrum fratrukket både døde og censurerede i tidsrummet
Overlevelses-sandsynlighed:	Sandsynligheden for at overleve fra sidste overlevelsestid til denne overlevelsestid. Dette bestemmes som andelen af overlevende ud af risikosættet.
Samlet overlevelses-sandsynlighed:	Sandsynligheden for at overleve et tidsrum som svarer til overlevelsestiden ind til nu. Dette bestemmes som produktet af alle overlevelsessandsynligheder indtil dette tidspunkt (da de enkelte overlevelsessandsynligheder er uafhængige)

Datasættet som I anvender og som bruges her indeholder 73 patienter. Vi vil nu lave en tabel der indeholder kolonner med ovenstående information.

Selve tabellen kan ses på næste side og er fremkommet bl.a. ud fra følgende argumentation: I datasættet er der 73 patienter, hvilket betyder, at risikosættet til start er 73 (der er 73 som kan dø). Sandsynligheden for at overleve ved dette tidspunkt (start) er selvfølgelig 1 (eller 100%), da patienterne var i live, da de fik diagnosen.

Ved andet tidspunkt (0,82 måneder) dør en person, dvs. at sandsynligheden for at dø (i tidsrummet fra 0 til 0,82) her er 1/73, eller tilsvarende er sandsynligheden for at overleve er 1-1/73 = 72/73.

Da der er en død ved andet tidspunkt og ingen censurerede op til næste (døds)tidspunkt vil antallet af personer som nu er i risikosættet være 72.

Den samlede overlevelsessandsynlighed fra start til dette dødstidspunkt bestemmes ved at gange sandsynligheden for at dø inden dette dødstidspunkt med sandsynligheden for at dø ved dette dødstidspunkt (da det er uafhængige hændelser).

På tilsvarende måde kan man bestemme risikosættets størrelse, overlevelsessandsynligheder (mellem to tidspunkter) og den samlede overlevelsessandsynlighed, hvilket ses af følgende:

Overlevelse (måneder)	Døde	Censurerede (inden næste dødstidspunkt)	Risikosæt	Overlevelsessandsynlighed (fra sidste overlevelsestid)	Samlet overlevelses-sandsynlighed
0	0	0	73	-	1
0,82	1	0	73	$1 - \frac{1}{73} = \frac{72}{73} \approx 0{,}9863$	$1 \cdot \frac{72}{73} \approx 0{,}9863$
1,02	1	1	$73 - (1 + 0) = 72$	$1 - \frac{1}{72} = \frac{71}{72} \approx 0{,}9861$	$0{,}9863 \cdot \frac{71}{72} \approx 0{,}9726$
2,96	1	0	$72 - (1 + 1) = 70$	$1 - \frac{1}{70} = \frac{69}{70} \approx 0{,}9857$	$0{,}9726 \cdot \frac{69}{70} \approx 0{,}9587$
2,99	1	5	$70 - (1 + 0) = 69$	$1 - \frac{1}{69} = \frac{68}{69} \approx 0{,}9855$	
11,63	1	0	$69 - (1 + 5) = 63$		
16,62	1	0			

1. Udfyld de grå felter i tabellen ovenfor
2. Hent regnearket "Data_til_KM_kurver.xlsx" og beregn samtlige overlevelsessandsynligheder (enten i Excel eller et andet sted) for både patienterne med ændringer i *BRCA1* og/eller *BRCA2* og for patienter uden.
3. Tegn begge kurver i samme diagram. For at få helt rigtige KM-kurver kan man med fordel anvende regnearket "Overlevelse_til _KM_kurver.xlsm" som kan ordne jeres beregnede data til brug for KM-grafer. Husk at gemme filen på din egen computer.
4. Ser det ud som om, der er forskelle på kurverne?

log-rank test - er en evt. forskel signifikant?

Vi vil nu se på, om forskellene vi så på KM-kurverne for de to grupper er signifikante, hvilket gøres ved at vurdere (statistisk) om de to kurver kunne stamme/komme fra samme kurve (eller om det er tilfældigt, at der er forskel). For at afgøre om dette er tilfældet, laver man beregninger på alle data (Både dem med og dem uden ændringer i *BRCA1* og/eller *BRCA2*) under hypotesen.

De to KM-kurver er ikke forskellige (dvs. sandsynligheden for på et givent tidspunkt at dø er ens i de to grupper)

Vi skal ud fra data bestemme, hvor mange forventede døde der er i hver af de to grupper (under hypotesen) for til sidst at afgøre, om de forventede antal af dødsfald i (hver af) de to grupper er væsentligt forskellige fra de observerede.

Hvis der som udgangspunkt ikke er forskel på de to kurver, er sandsynligheden for af en enkelt patient dør på et givent tidspunkt ikke afhængigt af, hvilken gruppe patienten tilhører. Antallet af døde i hver gruppe vil på denne måde kun afhænge af gruppernes størrelser (hvor mange i hver gruppe der kan dø (risikosættet)).

For at finde antal forventede døde er vi til hvert dødstidspunkt nødt til at kende risikosættets størrelse, hvilket igen betyder, at vi skal have styr på det observerede antal af døde og censurerede.

Risikosættenes størrelse til et bestemt tidspunkt bestemmes ved at fratrække antal døde og censurerede fra sidste tidspunkt fra risikosættet til sidste tidspunkt. Dette sker særskilt for gruppen med ændringer i *BRCA1* og/eller *BRCA2* og for gruppen uden ændringer *BRCA1* og/eller *BRCA2*.

Risikosættene markeret med blåt er bestemt ved, at risikosættene var henholdsvis 235 og 71 i 10. række. I 10. række dør en fra gruppen, hvor der ikke er ændringer i *BRCA1* eller *BRCA2*, hvilket betyder, at der bliver en mindre i dette risikosæt, mens det andet risikosæt bevares (række 11).

Efterfølgende kan antallet af forventede døde i hver gruppe bestemmes som sandsynligheden for at dø (under hypotesen) gange antallet af patienter i gruppen (risikosættet). Bemærk, at det forventede antal døde ved censureringer selvfølgelig er 0 for begge grupper, da ingen dør (til dette tidspunkt). Nedenfor er vist et udsnit af resultaterne, hvor udregningerne i de blå felter er vist efter tabellen, dvs. på næste side:

Række	Overlevelsestid (mdr.)	Antal døde		Antal censurerede		Risikosæt		Forventet antal døde	
		Ikke BRCA1 og BRCA2	BRCA1 og BRCA2	Ikke BRCA1 og BRCA2	BRCA1 og BRCA2	Ikke BRCA1 og BRCA2	BRCA1 og BRCA2	Ikke BRCA1 og BRCA2	BRCA1 og BRCA2
1	0,30	1	0	0	0	242	73	0,7683	0,2317
2	0,36	1	0	0	0	241	73		
3	0,76	1	0	0	0	240	73		
4	0,79	1	0	0	0	239	73		
5	0,82	0	1	0	0	238	73	0,7653	0,2347
6	1,02	0	1	0	0	238	72	0,7677	0,2323
7	1,18	1	0	0	0	238	71	0	0
8	1,18	0	0	1	0	237	71	0,7695	0,2305
9	1,97	0	0	1	0	236	71	0	0
10	2,00	1	0	0	0	235	71	0,7680	0,2320
11	2,14	1	0	0	0	234	71	0,7672	0,2328
12	2,17	0	0	0	1	233	71	0	0
13	2,23	1	0	0	0	233	70	0,7690	0,2310
14	2,46	1	0	0	0	232	70	0,7682	0,2318

309	116,04	0	0	0	1	2	2	0	0
310	118,99	1	0	0	0	2	1	0,6667	0,3333
311	125,63	0	0	1	0	1	1	0	0
312	180,04	0	0	0	1	0	1	0	0

Sum	146	35	96	38	-	-	124,3127	56,6873

Kort beskrevet er antallet af personer i risikogruppen for en given række bestemt af antallet af personer i rækken før fratrukket personer, der forsvinder (dør eller censureres) ligeledes i rækken før.

Det forventede antal døde bestemmes som sandsynligheden for at dø (under hypotesen), gange antallet af individer i gruppen. Sandsynligheden for at dø bestemmes som antallet af døde (fra begge grupper) divideret med antallet af personer fra begge risikosæt. Sandsynligheden bestemmes ud fra begge grupper, fordi den under hypotesen er antaget at være ens for begge grupper.

De blå felter under forventede antal døde beregnes som ud fra ovenstående som hhv.

$$\frac{1+0}{242+73} \cdot 242 \approx 0{,}7683 \quad \text{og} \quad \frac{1+0}{242+73} \cdot 73 \approx 0{,}2317$$

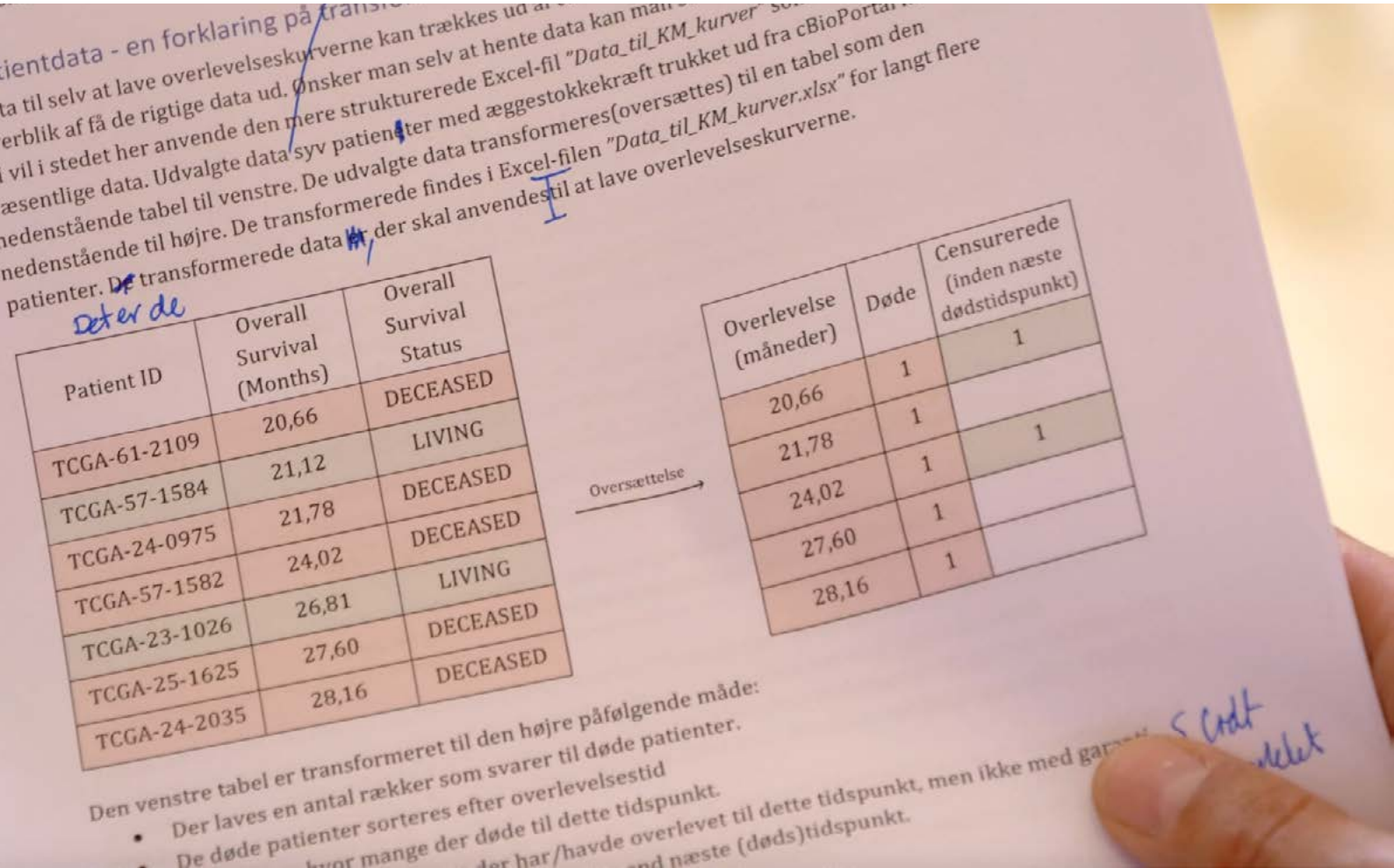
5. Bestem det forventede antal døde i begge grupper for anden, tredje og fjerde række (De grå felter). Skriv resultater ind i felterne.
6. Hent regnearket "Data_til_log-rank_test.xlsx" (husk at gemme filen på din egen computer) og beregn antal forventede døde for både patienterne med ændringer i *BRCA1* og/eller *BRCA2* og for patienter uden ændringer i *BRCA1* og/eller *BRCA2* for alle "dødstidspunkter" (enten i Excel eller et andet sted).
7. Vis at det samlede forventede antal døde i de to grupper er henholdsvis 124,31 og 56,69. (Tallet 56,69 er bestemt som summen af antal forventede døde i gruppen med ændringer i *BRCA1* og/eller *BRCA2* til hvert tidspunkt i tabellen).

Som det kan ses og som tidligere nævnt, har vi nu et sæt af to observerede værdier (antallet af døde i de to grupper) og to forventede værdier (det forventede antal døde i de to grupper) som kan sammenlignes med henblik på at vurdere, om kurverne er forskellige. Til dette formål bestemmes χ^2-teststørrelsen ved formlen

$$X^2 = \frac{(\text{obs}_1 - \text{forv}_1)^2}{\text{forv}_1} + \frac{(\text{obs}_2 - \text{forv}_2)^2}{\text{forv}_2}$$

Den fundne teststørrelse sammenlignes efterfølgende med en χ^2-fordeling med 1 frihedsgrad (1 frihedsgrad fordi antallet af frihedsgrader for en χ^2 GOF test bestemmes som antallet af grupper minus 1).

8. Bestem χ^2-teststørrelsen ud fra tabellen.
9. Bestem *p*-værdien ud fra teststørrelsen.
10. Kan vi afgøre, om der er signifikant forskel på de to KM-kurver ud fra teststørrelsen eller *p*-værdien?
11. Kan man afgøre, hvor stor forskellen er på de to KM-kurver ud fra ovenstående?



ientdata - en forklaring på transf

ta til selv at lave overlevelseskurverne kan trækkes ud af
erblik af få de rigtige data ud. Ønsker man selv at hente data kan man
i vil i stedet her anvende den mere strukturerede Excel-fil *"Data_til_KM_kurver*
æsentlige data. Udvalgte data syv patienter med æggestokkekræft trukket ud fra cBioPortal
nedenstående tabel til venstre. De udvalgte data transformeres(oversættes) til en tabel som den
nedenstående til højre. De transformerede findes i Excel-filen *"Data_til_KM_kurver.xlsx"* for langt flere
patienter. De transformerede data er der skal anvendes til at lave overlevelseskurverne.

Det er de

Patient ID	Overall Survival (Months)	Overall Survival Status
TCGA-61-2109	20,66	DECEASED
TCGA-57-1584	21,12	LIVING
TCGA-24-0975	21,78	DECEASED
TCGA-57-1582	24,02	DECEASED
TCGA-23-1026	26,81	LIVING
TCGA-25-1625	27,60	DECEASED
TCGA-24-2035	28,16	DECEASED

Oversættelse →

Overlevelse (måneder)	Døde	Censurerede (inden næste dødstidspunkt)
20,66	1	1
21,78	1	
24,02	1	1
27,60	1	
28,16	1	

Den venstre tabel er transformeret til den højre påfølgende måde:

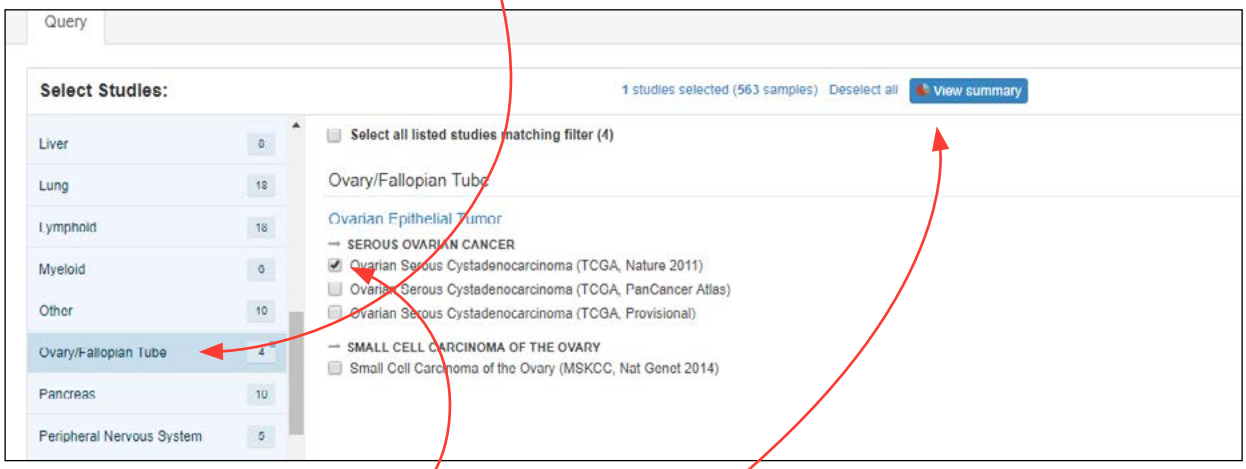
- Der laves en antal rækker som svarer til døde patienter.
- De døde patienter sorteres efter overlevelsestid
- hvor mange der døde til dette tidspunkt.
- der har/havde overlevet til dette tidspunkt, men ikke med gar
- end næste (døds)tidspunkt.

Sidst

Bilag 1: Hente data fra cBioPortal til brug for KM-kurver

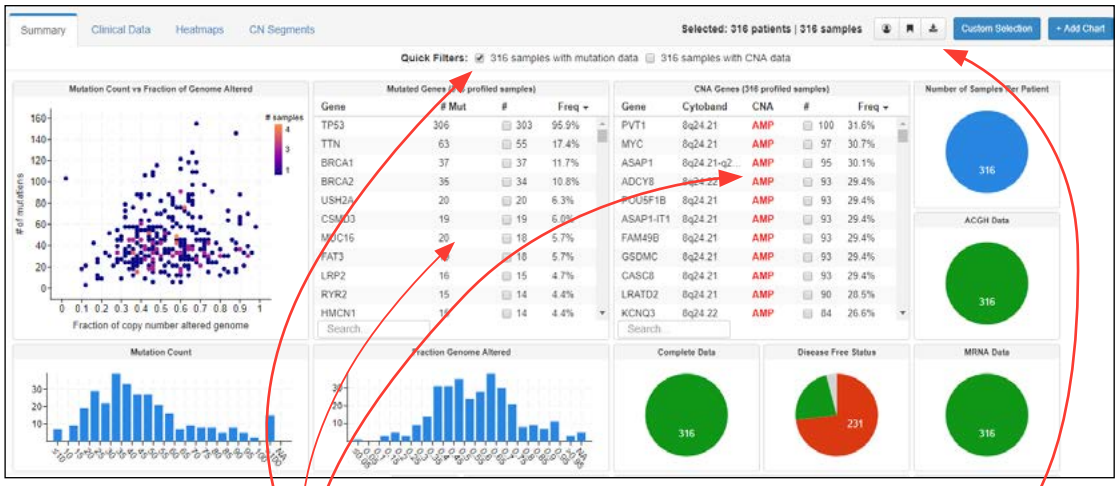
A. Gå til http://wwwcbioportal.org/

B. Under "Select Studies" findes og vælges "Ovary/fallopian Tube".



C. Der vælges "Ovarian Serous Cystadenocarcinoma (TCGA, Nature 2011)

D. Der trykkes på view summary



E. Her kan man så vælge data på forskellig vis og efterfølgende downloade valgte patient-data.

F. Filen der downloades har endelsen tsv, men er en tekstfil. Det kan dog være en fordel at ændre endelsen til txt, for at de fleste programmer opfatter, at det er en tekstfil (herunder Excel).

KØBENHAVNS UNIVERSITET
NØRREGADE 10
1165 KØBENHAVN K

WWW.KU.DK

KØBENHAVNS UNIVERSITET

Roskilde Gymnasium

novo nordisk fonden