Next Generation Sequencing

For 60 år siden blev DNA opdaget. I øjeblikket afprøver vi NGS, og til sommer kører vi BRCA-gener med den nye metode

I år er det 60 år siden, at DNA's struktur blev erkendt, og man fik en forståelse for, hvad DNA består af, og at DNA er vores arvemasse. Samtidig er det 10 år siden, at resultatet af den første totale sekventering af det humane genom forelå. I denne artikel vil vi fortælle om Next Generation Sequencing (NGS), som er en metode, der bruges til at sekventere, dvs. bestemme den nøjagtige rækkefølge af baserne i vores arvemasse. Sekventering kan bruges til at finde mutationer (mulig sygdomsfremkaldende variant) i vores arvemasse, DNA'et ved arvelige sygdomme. I vores afdeling, Klinisk Genetisk Afdeling på Odense Universitetshospital, bruges sekventering bl.a. til dette formål. Man kan også sekventere i mange andre situationer og på alle celler og arter.

Helt genom på én gang

Den nye teknik, NGS, har givet helt nye perspektiver, idet man med denne metode har mulighed for at få kæmpestore datamængder på relativt kort tid. Det er nye muligheder, som har stor betydning for forståelsen af biologien på planter, dyr og ikke mindst mennesker. Man kan nu sekventere hele genomet på én gang, det er stadig dyrt, men overkommeligt. Man kan sekventere hele exomet, dvs. at man sekventerer alle kodende områder af alle gener, eller man kan sekventere udvalgte gener eller områder af speciel interesse. Også for mRNA, microRNA eller regioner, der binder transkriptionsfaktorer, er der store muligheder med NGS.

Nye NGS-systemer udvikles

Indtil for få år siden blev al sekventering foretaget med Sanger-sekventering, som er en enzymatisk dideoxy-teknik, først beskrevet i 1977 [1]. I 1996 kom der automatiserede kapilær-instrumenter med fluorescerende dideoxy-nucleotider og laserdetektion. Pga. nytænkningen inden for sekventering blev NGS i 2007 valgt som årets metode af Nature Methods [2].

Fra 2005 blev der udviklet forskellige typer af instrumenter til NGS fra mange af de store firmaer, som beskæftiger sig med sekventering. Blandt flere kan man nævne Roche 454-systemet, som var det første succesfulde NGS-system. Pyrosekventeringsteknologi benyttes i dette system. SOLiD blev udviklet i 2006 af Applied Biosystems, som nu også har et nyere system, Ion Torrent (Life Tecnologies, som firmaet hedder nu).

Illumina kom til med Solexa GA og videreudviklede systemet til HiSeq 1000 og 2000. HiSeq 1000 er det instrument, som vores afdeling investerede i for et par år siden. Vores instrument er netop nu opgraderet til et HiSeq 1500-instrument, som har lidt flere muligheder end HiSeq 1000. Illumina udviklede desuden et mindre instrument, MiSeq, som vi også har på vores afdeling. MiSeq anvender en flowcelle med 1 lane i modsætning til HiSeq, som anvender flowceller med 8 lanes og er brugbart til store serier af prøver. Dog er det med opgraderingen til HiSeq 1500 muligt at køre 2 lanes, og instrumentet analyserer i kortere tid, hvilket er mere anvendeligt til klinisk brug.

Meget stor ydeevne

HiSeq 1000 (1500) leverer et meget stort output af sekventeringsdata. I tabel 1 ses datamængde, kørselstid og læselængde pr. kørsel på de instrumenter, som vi arbejder med.

Almindeligvis får man 300 x 10^9 brugbare baser ved en kørsel på HiSeq. Instrumentet har en kvalitetsscore, Q-score, som skal være >30 for 80 % af sekventerede baser (ved 2 x 100 bp). Vi tilstræber at læse hver base hundrede gange, dog mindst 30 gange til brug i et klinisk svar. Dette kommer vi nærmere ind på under metode og resultatvurdering.

Hvis man laver en anden sammenligning ud fra tabellen, kan man sige, at hele det humane genom er 300 x 10^9 baser, hvor instrumentet altså kan læse 300 x 10^9 baser eller 100 gange hele vores arvemasse.

Som man sikkert kan forstå, fylder denne store datamængde

Af afdelingsbioanalytiker Marianne Käehne, bioanalytiker Dorte Jensen og bioanalytiker Pernille Jordan // Klinisk Genetisk Afdeling, Odense Universitetshospital

	HiSeq 1000	HiSeq 1500	HiSeq 1500	MiSeq
Kørselsmetode	Højt output	Højt output	Hurtig kørsel	Højt output
Output (ved 2 x 100bp)	300 x 10^9 baser	300 x 10^9 baser	60 x 10^9 baser	3-3,4 x 10^9 baser
Kørselstid på instrumentet (2 x 100 bp)	8,5 dage	8,5 dage	27 timer	19 timer
Mulig læselængde	2 x 100 bp	2 x 100 bp	2 x 150 bp	2 x 250 bp (7,5 GB, 39 timer)
Sanger-sekventering, ABI 3730xl	Med vores nuværende metode læses op til 96 x 800 (i alt altså maks. ca. 75.000 baser i én kørsel, kørselstid ca. 1 time på instrumentet)			

TABEL 1.

rigtig meget på en computer, så man skal have stor regneplads og lagerplads til beregning og opbevaring af data.

Selve metoden

Prøveforberedelsen til kørsel på Illuminas HiSeq/MiSeq er tidskrævende, dog afhængigt af valgte metode. Den består af mange step og kan tage op til en uge. Herefter kan en kørsel/sekventering på HiSeq tage over en uge, hvis man kører med 8 lanes i flowcellen.

Imellem næsten alle step i prøveforberedelsen oprenses (med beads) og måles (Bioanalyzer/Qubit-fluoremeter) på produkterne for at sikre, at der arbejdes videre med rette størrelse på produktet, og at der er nok af det.

Først skal DNA-prøven (1 ug – 3 µg) oprenses fra EDTA-blod og fragmenteres til en på forhånd bestemt størrelse. Det vil sige, at DNA'et bliver "skåret" i små dobbeltstrengede stykker på ca. 150-200 bp. Dette foregår ved hjælp af lydbølger med en frekvens på 1 MHertz (på et instrument, som hedder Covaris). Tiden på Covaris varieres efter, hvor store stykker DNA som ønskes. Jo kortere tid, jo længere stykker. Hvor store stykker der ønskes, afhænger af, om det er hele genomet, hele exomet eller nogle bestemte gener, der skal sekventeres. Størrelse og kvantitet af det fragmenterede DNA måles herefter på en Bioanalyzer.

Adaptorer påsættes

Forskellige procedurer fører til, at det fragmenterede dobbeltstrengede DNA får repareret de iturevne ender, som er opstået ved fragmenteringen. Ydermere bliver der tilføjet et A- nucleotid, såkaldte adaptorer ligeres (sættes på) til begge ender af det dobbeltstrengede DNA-fragment (fig. 3). Disse adaptorer indeholder bl.a. en slags barcode, som består af 6 nucleotider, som er unikke for hver prøve. Dette muliggør, at prøverne kan blandes sammen, fordi de kan skelnes fra hinanden ved hjælp af sekvensen af disse barcoder. Jo flere prøver der blandes sammen, desto mindre data bliver opsamlet pr. prøve. Så hvis der kun køres en enkelt prøve, bliver der opsamlet rigtig meget data.

Kvantitering af den DNA-mængde, som påsættes flowcellen, er meget vigtig. Denne kvantitering foretages på et realtime-instrument, hvor man laver kvantitativ PCR ved hjælp af et kit, KAPA SYBR FAST ABI Prism qPCR Kit. Derefter sættes prøven på et instrument, C-bot, som ligeledes er fra Illumina.

I C-bot bindes DNA'et (milioner af små DNA-stykker) til

FIG 1.
HiSeq 1500, flowcellen ses i røret med orange låg.

FIG 2.
MiSeq, flowcellen med 1 lane ses i det lille rør med sort låg.

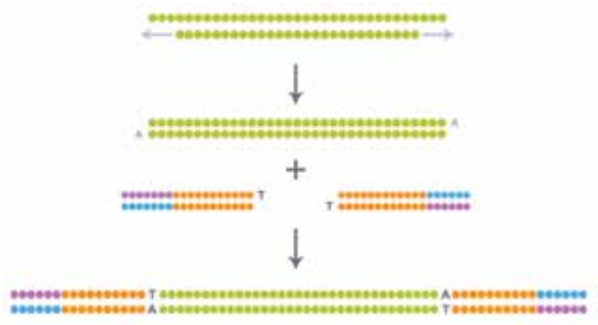

FIG. 3 viser procedure under sample preparation, "end repair", A ende og ligering af adaptorer til det færdige DNA-stykke, som kan bindes til flowcellen.

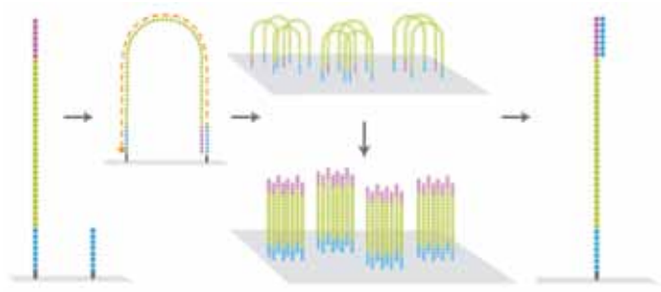

FIG. 4 viser dannelsen af clustre på flowcellens overflade.

flowcellen, idet sekvenser i ovennævnte adaptorer passer på tilsvarende oligonucleotider, som sidder på overfladen af alle lanes i flowcellen.

Kopier fremstilles

Desuden fremstilles en masse kopier vha. af en såkaldt "bridge amplification", se fig. 4. Denne "bro-opformering" gentages flere gange ved hjælp af reagenser i det fabriksfremstillede reagensrack, som købes til C-bot. Dette betyder, at man får dannet tusindvis af ens kopier af hvert lille stykke DNA, som er bundet til flowcellen. Disse fragmenter udgør et såkaldt cluster. Flowcellen kan indeholde 800.000 enkeltmolekyle- clusters pr. mm2. Hele processen i C-bot-instrumentet tager ca. 4-6 timer, og herefter er flowcellen klar til den endelige sekventering i HiSeq-instrumentet ved en teknik, som kaldes SBS-sekventering.

Et nucleotid pr. cycle

SBS betyder Sequencing by synthesis, altså sekventering via syntese. Til denne teknologi bruges 4 nucleotider (dNTP) med blokker. Nucleotiderne er fluorescensmærkede, og de indbygges en ad gangen svarende til den base, der sidder på DNA-stykket, som man skal sekventere. Eftersom der er mange ens små stykker DNA, indbygges de mærkede nucleotider parallelt på alle DNA-stykker, fluorescensen registreres, og blokker og fluorescensen afskæres enzymatisk (1 cycle), hvorefter der indbygges et nyt nucleotid osv. For hver cycle indbygges altså 1 nucleotid, og man skal således køre 100 cycles for at læse en længde på 100 baser på en DNA streng. Reagenserne til sekventeringen pumpes igennem flowcellen under kørslen, og en scanner registrerer fluorescensen. Se fig. 5.

Databehandling

Efter kørslen skal den meget store datamængde behandles. Vi har 2 bioinformatikere ansat til at opsætte systemer til databehandling, men på sigt skulle bioanalytikerne gerne selv kunne foretage en stor del af databehandling. Der benyttes en del forskellige softwareprogrammer, bl.a. CASAVA-, GATK- og NOVOALIGN-software, hvor de sekventerede baser sammenlignes med en referencesekvens (fra genom-databaser). En sådan sammenligning med en database foretages også på nuværende tidspunkt, så dette er der ikke den store forskel i, blot at datamængden ved NGS er så kæmpestor i forhold til traditionel sekventering.

De kriterier, som vi arbejder efter, er, at vi skal have en tilstrækkelig dækning over hele genet, alle kodende områder. Ved dækning forstås, at alle baser og alle områder skal læses et vist antal gange, for at vi kan være sikre på resultatet. Som minimum skal alle kodende områder, exons, have en dækning på mindst 30, altså være læst mindst 30 gange. Af fig. 6 ses, at BRCA-genet i denne kørsel læses over 30 gange i alle exons.

Resultater

De resultater, som vi har indtil nu, er følgende:

Alle tidligere undersøgte varianter (Sanger-sekventering) har vi kunnet genfinde på den ny metode, NGS. I dette materiale findes både enkeltbasesubstitutioner, stopmutationer, splicemutationer samt små og store deletioner og insertioner. Det næste skridt er således parallelt at køre Sanger-sekventering og NGS på de patientprøver, som vi får ind til analysering.

Vi forventer at være klar til klinisk brug inden sommeren 2013. Herefter vil der stadig foregå en del optimering, ikke mindst fordi NGS er i en rivende udvikling, og både Illumina, Agilent og andre firmaer, som udvikler og sælger reagenser, haster derudad med udvikling af nye, smartere og hurtigere metoder. Vi kan knap nå at afprøve én metode, før den er videreudviklet.

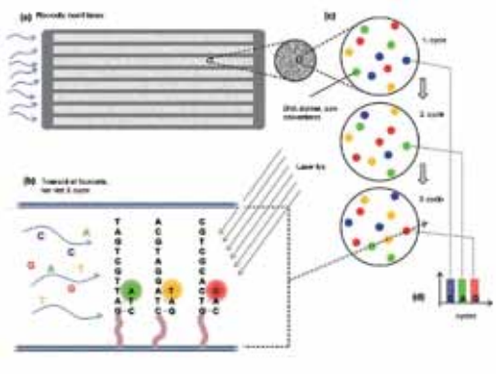

FIG. 5 viser teknikken SBS-sekventering, altså sekventering via syntese.

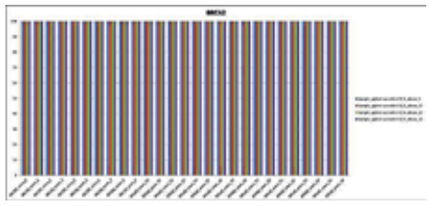

FIG. 6 viser, at over alle kodende områder i BRCA2-genet er dækningen over 30.

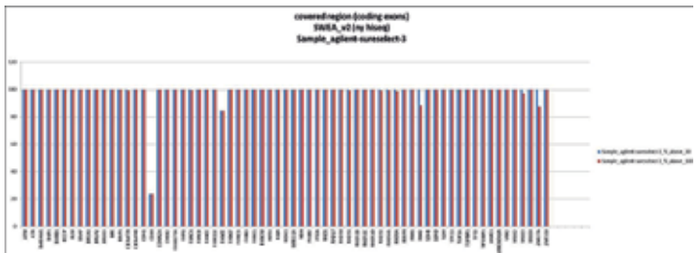

FIG. 7 viser dækningen over alle gener i vores "cancerpakke". Som man ser, er ikke alle gener dækket lige godt på nuværende tidspunkt. Disse skal om muligt optimeres, hvis vi vil afgive svar på disse.

Afprøvningen i vores laboratorium

I indkøringsfasen af NGS på HiSeq 1000 (1500) har vi afprøvet forskellige metoder, dog er alle metoder kørt med et sample preparation kit fra Illumina.

Vi har afprøvet Truseq exom enrichment kit fra Illumina, her sekventeres alle kodende områder i generne i hele genomet.

Vi har også afprøvet custom Truseq enrichment kit ligeledes fra Illumina. Det er en speciel designet "fiskestang" til 70 gener i laboratoriet, hvor både intron og exon sekventeres.

Af lidt mere specielle metoder, som vi har afprøvet, er først at køre sample preparation Truseq kit fra Illumina efterfulgt af SureSelect fra Agilent. Altså to protokoller fra forskellige firmaer, der er sat sammen. Her er det også en speciel designet "fiskestang", hvor vi ser på både intron og exon. Dette er afprøvet både på BRCA og på 21 andre gener, som vi undersøger i laboratoriet i dag.

Endelig har vi afprøvet SureSelect Target Enrichment System fra Agilent, hvor vi har brugt SureSelect-protokollen hele proceduren igennem.

Starter med BRCA-gener

Vores strategi er at starte med at sekventere BRCA-generne i en pakke på i alt 64 gener og analysere ved hjælp af SBS-systemet på HiSeq 1500. Dette er vi klar med forår/sommer 2013. Dvs. at vi i første omgang kun vil afgive svar på BRCA1 og BRCA2, selvom der er kørt yderligere 62 gener, som har med cancer at gøre. Forventningen er selvfølgelig også, at vi, alt efter hvilke ønsker vores brugere måtte have, kan afgive svar og udføre rådgivning på de øvrige gener i pakken.

Senere følger alle andre gener, som vi undersøger i laboratoriet i dag. Her har vi dog ikke afgjort, hvilken metode vi vil anvende.

Verificeringen af fundne mutationer ved hjælp af SBS-metoden skal stadig foregå ved Sanger-sekventeringsmetoden. Ligeledes skal kendte mutationer (dvs. mutationer, som kendes i forvejen i familien) også sekventeres ved Sanger-metoden, som vi gør i dag, da SBS-metoden er for dyr og for tidskrævende til dette formål.

Vores plan er desuden, at alle de fund, som findes ved NGS på screening af gener, i første omgang BRCA1 og 2, skal efterprøves med traditionel sekventering på en kontrolprøve fra samme patient.

Etik og NGS

Ved hjælp af denne nye teknik – NGS – kan man kortlægge det humane gen på kort tid. Den gør det nemlig muligt at læse én persons ca. 25.000 gener inden for få dage på én enkelt maskine.

Teknikken gør det derved muligt at finde frem til sjældne arvelige sygdomme langt hurtigere end tidligere, og man kan derved afkorte tiden, det tager at stille diagnosen, så den rigtige behandling kan iværksættes.

Men fordi metoden afslører stort set alle genforandringer, også i gener, der måske ikke er relevante for den aktuelle sygdom, der undersøges for, medfører NGS-metoden store etiske udfordringer. Teknikken kan afsløre genforandringer, som vi med vores viden på området i dag ikke kender betydningen af. Dette stiller store krav til videre forskning og ikke mindst til rådgivningen af patienter.

Vil man kende sandheden?

Spørgsmålet er så, om man er klar til at få kortlagt sit genom og få hele sandheden om sit helbred – også viden om alvorlige sygdomme, som med stor sandsynlighed vil ramme én uden at kunne behandles. Man kan få en viden om sit genom, som bliver svært at forholde sig til, også selv om det måske drejer sig om en let forøget sygdomsrisiko. Man bliver måske unødig bekymret, selvom man egentlig er rask.

Hvis man fx vælger at få kortlagt sit genom, dette er dog

ikke muligt via sundhedssystemet her i Danmark på nuværende tidspunkt, kunne det vise sig, at man er disponeret for en alvorlig sygdom, fx Huntingtons Chorea. Denne viden vil få stor betydning for ens tilværelse, og da sygdommen ofte først viser sig i 50-års alderen, får man et liv med meget store dilemmaer. Hvis man har sat børn i verden, inden man får kendskab til sin egen sygdomsrisiko, så ved man med denne sygdom, at ens børn ligeledes vil få sygdommen og overvejende sandsynligt på et tidligere tidspunkt end én selv. Hvis man omvendt endnu ikke har fået børn, vil man kunne afholde sig fra dette, hvilket vil være en god ide i relation til sygdomsspredning.

Hvad gør vi med "ekstra" viden

Når vi går i gang med NGS her på afdelingen, starter vi med at analysere BRCA-generne, men vi vil også få viden om 62 andre gener, som har med cancer at gøre. Skal vi vælge at kigge på dem alle og måske få en viden, der vil kunne afsløre risikoen for en række andre cancersygdomme, eller skal vi lægge et filter ind, så vi kun ser på BRCA-generne? Kan vi overhovedet rådgive patienterne på nuværende tidspunkt? Hvad vil vi vide? Hvad vil patienten vide, og hvad kan genetikerne fortælle patienten? Hvad stiller patienten op med den viden, som vi kan forudse én eller anden risiko for? Skal vi lukke øjnene og vælge kun at se på BRCA-generne, som er den sygdom, patienten egentlig skulle undersøges for, eller skal vi kigge mere bredt og derved sidde med en viden, som vi ikke ved, hvad vi skal gøre med? Ja, der er mange nye etiske udfordringer og spørgsmål, der opstår i forbindelse med denne nye fantastiske teknologi – NGS.

Endnu hurtigere metode på vej

Samtidig med denne fortsatte udvikling af både reagenser og instrumenter til NGS og den tiltagende brugbarhed af teknikken arbejdes der videre med det, man kan kalde third generation sequencing (PacBio RS og Nanopore). Til denne teknologi er der 2 hovedkarakteristika. Man laver ikke PCR før sekventering, og signalet opsamles i realtime, altså samtidig med sekventeringen. Dette nedsætter analysetiden væsentligt – men altså lad os lige indføre NGS, før vi kaster os over det næste store fremskridt. □

Referencer:

[1] Sanger, F. et al. (1977): DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A 74, 5463-5467

[2] Schuster, S.C. et al. (2008): Method of the year, next-generation DNA sequencing. Functional genomics and medical applications. Nat. Methods 5, 11-21

Ordliste:

Sekventere: at finde den helt præcise rækkefølge af baser i vores arvemasse

humane genom: hele vores arvemasse

Exomet: alle kodende områder i vores arvemasse, dvs. alt det arvemateriale, som koder for et eller andet protein, som kroppen skal bruge

Mutation: betydende fejl i vores arvemasse

Gen: del af arvemassen, som koder for et bestemt protein

Exon: kodende (for protein) område af et gen

Intron: ikke kodende område af et gen

Kapilær-instrumenter med fluorescerende dideoxy-nucleotider: sekventeringsmaskiner, som automatisk suger prøverne op og aflæser fluorescensen

Flowcelle: DNA'et sættes fast på flowcellen, og flowcellen gennemstrømmes af reagenser, således at sekventeringen kan foretages

Lane: "bane" i flowcellen

Læselængde: betyder i denne sammenhæng, hvor store DNA-stykker der kan sekventeres ad gangen.

Clustre: en masse kopier af samme stykke DNA i et meget lille område på flowcellens overflade

Nucleotid: bestanddel i DNA-molekylet

Dideoxy-nucleotid: et nucleotid, som mangler et iltmolekyle, som bruges til at viderebygge en DNA-streng (laboratoriefremstillet enhed, som ikke findes i mennesket), bruges kun analyseteknisk.

Cycle: 1 runde med reagenser

Genom-database: database, som rummer en reference-DNA-streng, som den skal være, for at man kan sige, at DNA'et er "normalt" (normalområde)

Kontrolprøve: I vores laboratorium udbeder vi os altid 2 blodprøver fra patienten. Hvis vi finder en mutation, efterprøver vi, om vi også kan finde den i blodprøve nr. 2. Primærprøve og kontrolprøve køres hver for sig, vi sikrer os derved mod forbytninger i laboratoriet.

PCR: bruges til at opformere DNA'et i prøven

Paired-end betyder, at man først læser 100 baser fra den ene ende af DNA-stykket, tilsætter nogle nye reagenser til instrumentet (HiSeq) og dernæst læser 100 baser fra den anden ende af DNA-stykket.